UNIVERSIDAD AUTÓNOMA DE MADRID

FACULTAD DE CIENCIAS

DEPARTAMENTO DE BIOLOGIA MOLECULAR

The specific roles of cohesin-SA1 and cohesin-SA2 in gene regulation and genome organization

Aleksandar Kojic

Degree in biology

Thesis Directors:

Dr. Ana Losada Valiente and Dr. Ana Cuadrado Garcia

Centro Nacional de Investigaciones Oncológicas

Madrid, 2018

Summary

The use of Chromosome Conformation Capture (3C)-derived techniques for probing 3D genome structure has shed new light on the principles of genome organization. DNA looping is recognized as the major mechanism responsible for this organization in which topologically associated domains (TADs) and compartments are acknowledged as hierarchical levels of genome folding at a megabase scale. The cohesin complex, working together with the insulator protein CTCF, has shown to be essential for partitioning the genome into chromatin loops and TADs. Cohesin is a ring-shaped protein complex best known for its role in sister chromatin cohesion that consists of four subunits: SMC1, SMC3, RAD21 and SA. In somatic vertebrate cells, the SA subunit can be either SA1 or SA2, thus giving rise to two different cohesin variants, cohesin-SA1 and cohesin-SA2. Studies in human and mouse cells have suggested that these variants have non-redundant functions, at least regarding cohesion. However, their differential contributions to gene regulation and genome organization have not been explored.

In order to address this question, we first analysed the genome-wide distribution of the two cohesin variants in three human cell lines and found that cohesin-SA1 is almost exclusively present at CTCF-bound sites. In contrast, only a fraction of cohesin-SA2 colocalizes with cohesin-SA1 and CTCF, while another one occupies non-CTCF positions that are instead bound by transcriptional regulators. This population is enriched at active enhancers, in particular at super-enhancer elements responsible for establishment of cell identity. Importantly, cohesin-SA2 can reach CTCF positions in the absence of SA1, but cohesin-SA1 cannot occupy non-CTCF positions when SA2 is absent. Differential affinity of SA1 and SA2 for cohesin-releasing factor WAPL, CTCF and transcriptional regulators may contribute to the observed differences in the dynamics of their association to chromatin. Downregulation of cohesin-SA2 has a more pronounced effect on the transcriptome of MCF10A cells and leads to deregulation of core cell-identity genes. Using 4C and Hi-C analyses we also explored the consequences of cohesin variant depletion on local and global genome organization. Our results suggest that cohesin-SA1 works together with CTCF to define TADs while cohesin-SA2 facilitates more transient and local regulatory contacts. In summary, we propose that cohesin-SA2 contributes to cell-type specific gene regulation in a CTCF-independent fashion, a function that cannot be assumed by cohesin-SA1. Our work provides a new perspective on understanding the contribution of cohesin mutations to the pathology of human cancers.

Resumen

El reciente desarrollo de técnicas de Captura de Conformación Cromosómica (3C) y su aplicación al estudio de la estructura 3D de la cromatina ha supuesto un gran avance en nuestro conocimiento de la organización genómica. La formación de "lazos" de DNA es el principal mecanismo responsable de dicha organización, en la que los TADs (dominios asociados topológicamente) y los compartimentos (A y B) constituyen diferentes niveles de plegamiento genómico. La cohesina y CTCF son piezas fundamentales en la formación de estas estructuras. La cohesina es un complejo proteico con forma de anillo, descubierto por su papel en la cohesión de cromátidas hermanas. Está formado por tres subunidades comunes, SMC1, SMC3, RAD21 y una cuarta, la proteína SA, que en vertebrados posee dos isoformas, SA1 y SA2. De este modo, en las células somáticas coexisten dos variantes de cohesina, cohesina-SA1 y cohesina-SA2. Diversos estudios en líneas celulares de mamíferos sugieren que ambos complejos poseen funciones no redundantes, al menos en relación a su participación en el proceso de cohesión. Quedaba sin embargo por explorar su contribución a la regulación de la organización genómica y la expresión génica, principal objetivo de esta Tesis.

El análisis de la distribución genómica de los dos complejos cohesina en tres líneas celulares humanas ha revelado que mientras que la cohesina-SA1 se encuentra exclusivamente en sitios unidos a CTCF, sólo una fracción de la cohesina-SA2 se localiza en este tipo de posiciones, en tanto que el resto se distribuye en regiones sin CTCF y que ocupadas en cambio por reguladores transcripcionales. Estas regiones se corresponden principalmente con "enhancers" activos, en particular, "super-enhancers", responsables de la transcripción que determina la identidad celular. En ausencia de SA1, cohesina-SA2 puede seguir ocupando las mismas posiciones con o sin CTCF que ocupa normalmente. Sin embargo, cohesina-SA1 no puede sustituir a la cohesina-SA2 en las regiones sin CTCF. Ambos complejos muestran diferencias tanto en su movilidad como en su estabilidad en la cromatina, diferencias que pueden estar relacionadas con su distinta afinidad por Wapl, CTCF u otros reguladores transcripcionales. La reducción de los niveles de SA2 tiene un mayor impacto en la transcripción, afectando a la expresión de genes críticos para la identidad celular. El análisis de la contribución específica de cada variante a la organización genómica mediante Hi-C nos permite concluir que mientras que cohesina-SA1 colabora con CTCF en la definición de los TADs, cohesina-SA2 está implicada en la formación de contactos reguladores, locales y transitorios. Por todo ello, proponemos que cohesina-SA2 contribuye a establecer los programas de expresión génica determinantes del tipo celular, función que no puede llevar a cabo cohesina-SA1. Este trabajo aporta una nueva perspectiva para entender los posibles mecanismos de patogenicidad de las mutaciones en cohesina identificadas en cáncer.

Table of contents

Table of Contents

Abbreviations

3C: Chromosome Conformation Capture

ADP: Adenosine diphosphate

ATP: Adenosine triphosphate

CdLS: Cornelia de Lange Syndrome

ChIP: Chromatin Immunoprecipitation

ChIP-qPCR: Chromatin Immunoprecipitation followed by quantitative PCR

ChIP-seq: Chromatin Immunoprecipitation coupled with deep sequencing

CHOPS: C for cognitive impairment and coarse facies, H for heart defects, O for obesity, P for pulmonary involvement and S for short stature and skeletal dysplasia

co-IP: co-Immunoprecipitation

CTCF: CCCTC-binding factor

DEG: Differentially Expressed Gene

dpc: days post coitum

Esco1/2: Establishment of Cohesion 1/2

FACS: Fluorescence activated cell sorting

FISH: Fluorescence In Situ Hybridization

FPKM: fragments per kilobase per million of reads

GSEA: Gene Set Enrichment Analysis

GWAS: Genome Wide Association Studies

HCAEC: Human Coronary Artery endothelial Cells

HDAC8: Histone Deacetylase 8

HEAT: Huntingtin, elongation factor 3 (EF3), protein phosphatase 2A (PP2A), and the yeast kinase TOR1

HMEC: Human Mammary Epithelial Cells

HUVEC: Human Umbilical Vein endothelial Cells

IFNG: Interferon Gamma

KO: knockout

MEFs: Mouse Embryonic Fibroblasts

NIPBL: Nipped-B-like protein

NGS: Next Generation Sequencing

n.s.: non significant

Pds5: Precocious dissociation of sisters 5

PP2A: Protein Phosphatase 2A

pre-RC: pre-Replicative Complex

PTM: post-translational modification

RAD21: Double-strand-break repair protein 21 homolog

RNA-seq: RNA-sequencing

RNApolII: RNA polymerase II

RBS: Roberts Syndrome

RT-qPCR: Retrotranscription followed by quantitative PCR

SE: Super-enhancer

SEC: super elongation complex

Sgo1: Shugoshin 1

siRNA: small interfering RNA

SMC: Structural Maintenance of Chromosomes

SA: Stromal Antigen (protein)

STAG: Stromal Antigen (gene)

TAD: Topologically Associated Domain

TSS: Transcription Start Site

WAPL: Wings apart-like

WB: Western Blot

wt: wild type

Introduction

1. Genome organization is important for its function

Deciphering the principles of three-dimensional (3D) organization of the genome has always been a distinctive objective in biology. The size of the DNA poses an incredible challenge to the cell: almost 2-meter long DNA has to be packed in the 10-μm diameter of the human cell nucleus. Moreover, a single cell -the zygote- has to give rise to a complex organism composed of multiple cell types, all of them using the same genetic information. The importance of three-dimensional chromatin organization lies in providing a manageable chromosome size but also in the fact that folding is not random and contributes to regulate genome function (Bonev and Cavalli, 2016).

How the genome folds appears to be particularly important for transcription since regulatory elements, such as enhancers, are often far away from their target genes. Moreover, c*is*-regulatory elements occupy a much larger fraction of the genome than the protein coding genes, with multiple enhancers acting on a single gene. Enhancer–promoter contacts must be tightly regulated in order to prevent spurious activation of non-target neighboring genes (de Laat and Duboule, 2013). 3D chromatin structure transitions have been observed in development and cell differentiation, and are often dysregulated in disease processes (Norton and Phillips-Cremins, 2017). Structural variations in the human genome such as deletions, inversions or translocations that alter genome topology have been shown to cause malformation syndromes (Lupiañez et al., 2015) and to promote tumourigenesis though oncogene activation (Hnisz et al., 2016). It has also become evident that most disease associated genomic variants identified in Genome Wide Association Studies (GWAS) are present in cis-regulatory elements. Indeed, in the current post-GWAS era, a challenging task is to go beyond traditional exome-sequencing approaches in order to identify non coding variants in enhancers as major candidates for functional interpretation of GWAS loci. Identification of the gene whose altered expression promotes the disease requires understanding of the contacts in which the mutated element is engaged (Smemo et al., 2014). Clearly, understanding genome architecture is key not only from the perspective of basic knowledge, but also for human disease.

Chromosomes fold in a hierarchy of structures with increasing complexity (Figure I1). At the bottom of the hierarchy is the nucleosome in which 147-bp long stretch of DNA wraps around histone octamers composed of two copies each of H2A, H2B, H3, and H4. The amino- and carboxy-terminal tails of histones are subject to several post-translational modifications (PTMs) such as methylation, acetylation, phosphorylation, sumoylation, ADP-

ribosylation, or ubiquitinylation (Taverna et al., 2007). Histone modifications determine how nucleosomes interact with each other, control access to given sequences, and recruit effector proteins that ultimately determine the activity of the region in terms of transcription, recombination, etc. As we will explain below, functional elements of the genome are characterized by the presence of certain combination of histone marks.

For a long time it was thought that the 10-nm chromatin fiber, composed of nucleosomes, would fold into solenoid-like arrays to form a 30nm chromatin fiber, but the existence of this entity *in vivo* is unclear (Fussner et al., 2012). Instead, nucleosomes appear to be arranged in heterogeneous groups, called 'clutches', in a cell-type dependent manner (Ricci et al., 2015).

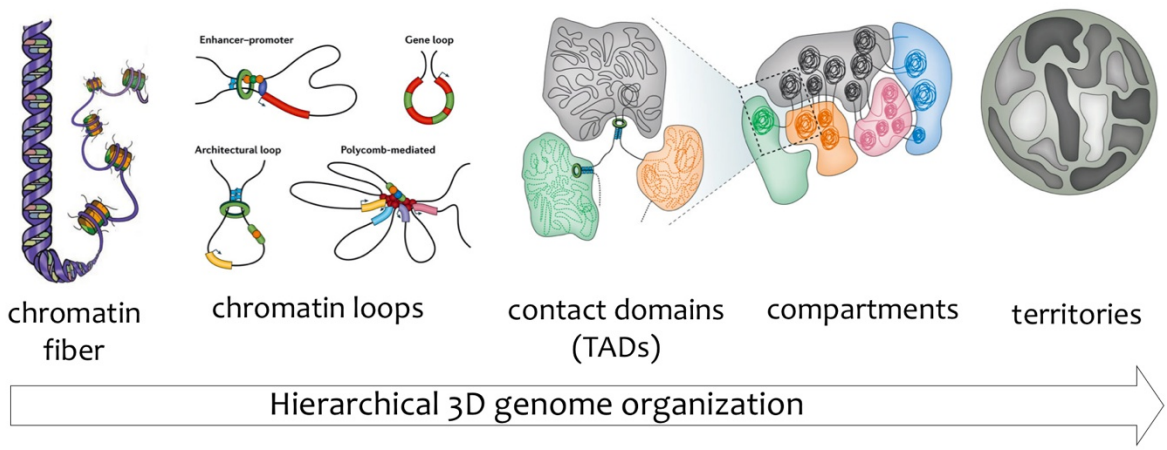


Figure I1. Hierarchy of genome organization

Chromosomes fold in a hierarchy of structures with increasing complexity, from nucleosomal DNA to non-random arrangements of individual chromosomes (territories) within the nuclear space. Adapted from Bonev and Cavalli, 2016.

On the upper end of the hierarchical organization of genomes we find individual chromosome territories, as observed by fluorescence in situ hybridization (FISH, Cremer and Cremer, 2001). Between the chromatin fiber and the chromosome territories, genome folding has been a black box for many years. A large number of contributions in the last two decades have driven dramatic and rapid advances in the field of 3D genome structure and have provided a better insight into the link between spatial organization of genome and its function. Our current view of genome organization is based largely on data obtained with two experimental approaches: FISH to visualize directly the proximity between two or more DNA segments in individual cells and chromosome conformation capture (3C) and its derivatives, which assess contact frequencies between selected genomic sites in cell populations. I will

briefly review these methodologies in section 1.1 before continuing with the description of the higher levels of chromatin organization in section 1.2.

1.1.Techniques for probing genome organization

The development of Chromosome Conformation Capture (3C) and its derived methods to probe genomic contacts between distant loci has been a major breakthrough in the field of genome organization (Dekker et al., 2002; de Wit and de Laat, 2012). This technique relies on the idea that digestion and re-ligation of fixed chromatin in cells, followed by quantification of the ligation junctions, reveals DNA contact frequencies and thereby provides insight into chromosome topology (Figure I2).

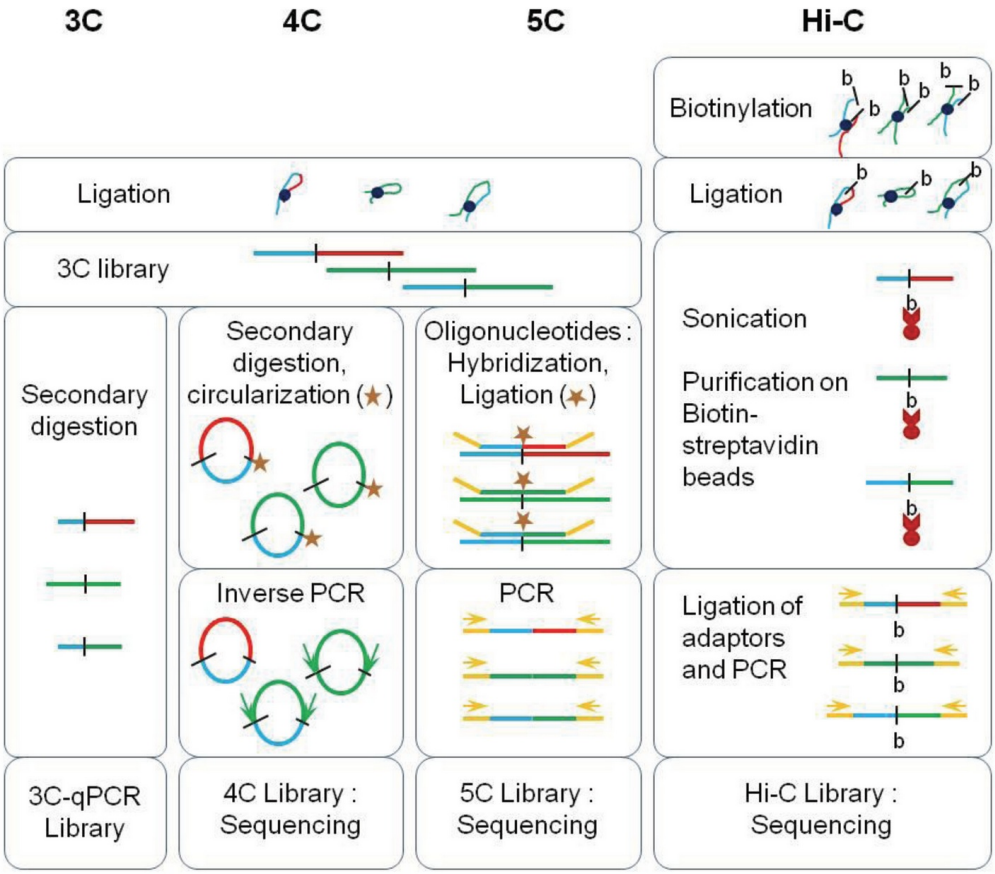


Figure I2. Chromosome Conformation Capture (3C) technologies

All protocols of Chromosome Conformation Capture (3C) involve a formaldehyde cross-linking step followed by an enzymatic digestion and a ligation step. Each method uses then different approaches to generate genomic libraries: secondary digestion for 3C-qPCR, circularization and inverse PCR for the 4C, “carbon copy” amplification for the 5C and biotinylation and purification on streptavidin beads for Hi-C. Ligation products are quantified by real-time qPCR in the first method or by high-throughput sequencing in the others. Adapted from Ea et al., 2015.

The original 3C technique was based on PCR detection of specific pair of fragments and it was therefore limited to detection of interactions between loci for which some previous knowledge existed. High-throughput versions of 3C, including “4C”, “5C” and “Hi-C”, along with the concomitant development of next-generation sequencing (NGS) technologies, allowed for generation of high-resolution maps of genome-wide contacts. Libraries are generated from millions of cells and thus represent a complex mixture of ligation events that reflect the average interaction profile of each fragment in the population across the whole genome. Single-cell Hi-C approaches have been recently developed and reveal the stochastic nature of the genome folding (Flyamer et al., 2017).

Even though every 3C-derived technique has its application depending on the biological question that should be answered, Hi-C is the technique of choice for taking an insight into genome-wide interaction profiles (Liberman-Aiden et al., 2009). In the Hi-C protocol the restriction ends are filled with biotinylated nucleotides before ligation. Following sonication, affinity purification on streptavidin beads ensures that only ligation junctions are selected for further analysis, greatly reducing the complexity of the mixture to be sequenced. Reads are mapped to the genome, and when a pair is found on two different restriction fragments, this is scored as an interaction between these two fragments. From this, a matrix of ligation frequencies between all fragments in the genome is generated. With increased sequencing depth, higher resolution can be achieved in the contact maps, although a 10-fold increase in resolution requires a 100-fold increase in sequence depth.

To look at selected loci in the genome at high resolution, the 4C technique can be used (Simmonis et al., 2006). In this case, a viewpoint is set in the region of interest and all the contacts of this fragment with the rest of the genome are detected. In 4C, the ligated 3C template is subject to a second restriction digest and ligation to generate small DNA circles, some of which will contain the 3C ligation junctions. Inverse PCR using viewpoint-specific primers is then used to amplify all sequences contacting this chromosomal site, which are then analyzed by NGS methods. We previously used 4C-seq to probe the architecture of two regions of the mouse genome in embryonic brains and adult pancreas (Cuadrado et al., 2015).

Apart from the application in 3C-derived assays, NGS technologies are used with Chromatin Immunoprecipitation (ChIP-seq) to generate genome-wide binding maps of DNA-binding proteins as well as genome-wide epigenomic profiles (Barski et al., 2007; Mikkelsen et al., 2007). Covalent histone modifications together with DNA methylation constitute the epigenome of the cell (Kouzarides, 2007). A number of consortia have used ChIP-seq as a

tool to provide a panel of histone modifications for each cell type in order to gain an insight into cell-type specific epigenomes (https://www.cell.com/consortium/IHEC?code=cell-site). This initiative has generated a large amount of data for which development of integrative computational approaches was necessary to capture the patterns of histone modifications and translate them into biologically meaningful genomic features called chromatin states (Ernst et al., 2011). Two popular tools are ChromHMM and Segway. In both approaches, the ChIP-seq experiments are transformed into genome-wide multivariate signals and subsequently used as observed variables in a probabilistic inference algorithm (Mammana and Chung, 2015).

In addition to 3C and NGS methods, the field of genome organization has benefited from the development of super-resolution microscopy (Boettiger et al., 2016; Fabre et al., 2015; Wani et al., 2016) as well as from the application of polymer and statistical physics models to simulate genome organization. These simulations can, based on the existing data coming from 3C-derived assays, predict future experiments or can work in the opposite direction and, based on simple biological assumptions, deliver output that can be compared to existing Hi-C contact maps (Haddad et al., 2017; Bianco et al., 2017).

Combination of outlined techniques led to paradigm shift in the understanding of function and higher-order organization of interphase genome, more specifically at scales of 10-100s of kilobases in size, corresponding to organizational levels between nucleosomes and chromosome territories.

1.2. Shedding light into the black box: from chromatin loops to compartments

Hi-C experiments revealed that the genome is divided into sub-chromosomal compartments named “A” and “B” (Lieberman-Aiden et al., 2009). Compartments represent accumulation of large (several Mb) chromosomal domains which are usually located on the same chromosome, giving rise to “plaid” pattern visible on genome-wide interaction matrices obtained from Hi-C. They reflect tendency for regions of similar chromatin status and epigenomic type to be in a close proximity, separated from loci of opposite features (Lieberman-Aiden et al., 2009; Zhang et al., 2012). Active, gene-rich and DNAse I hypersensitive loci are preferentially clustered in A compartments whereas transcriptionally silent, gene-poor and DNAse I insensitive loci are converging to form B compartments. Due to their similarity, it is plausible that lamina-associated domains (LADs) and nucleolus-associated domains (NADs) are actually B-like accumulations of loci at nuclear lamina and nucleoli (Gibcus and Dekker, 2013) (Figures I1 and I3).

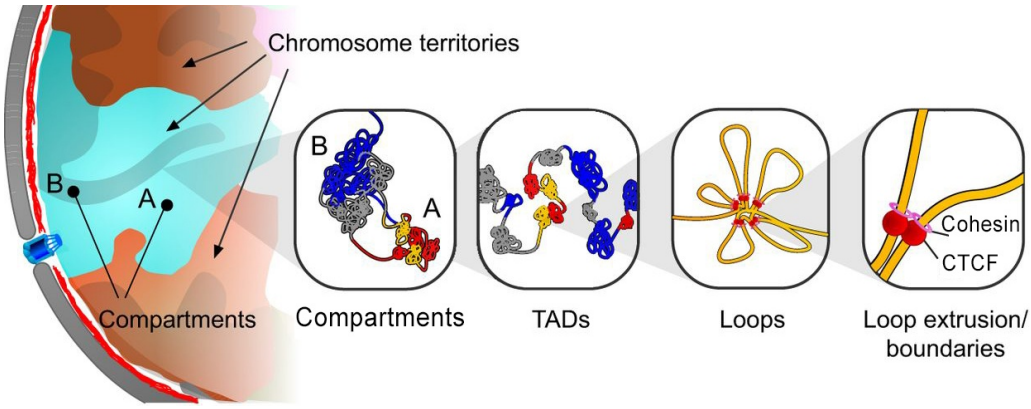


Figure I3. Chromatin domain folding at different scales

Chromosome territories can be split into A and B compartments, transcriptionally more active or inactive, respectively. The A/B compartments are organized in superdomains, which include mostly domains belonging to the same chromatin type. Superdomains group together topologically associated domains (TADs) of similar chromatin type. Inside TADs, loop formation is favored, in particular between enhancers and promoters. A number of proteins previously characterized as insulators show high enrichment between TADs and/or at the base of the loops. Loops are formed by the combined action of loop extrusion factors (SMC complexes) and boundary factors (such as CTCF) (Taken from Pueschel et al., 2016).

Subsequent Hi-C and 5C experiments reported the existence of smaller contact domains within larger A and B compartments named TADs for “topologically associated domain” (Nora et al., 2012; Dixon et al., 2012; Sexton et al., 2012). These sub-megabase domains are defined as regions with high propensity to self-interact, partially insulated from adjacent TADs. In other words, long-range interactions between loci within the same domain occur with higher frequency than interactions between loci located in adjacent domains (Figures I1 and I3).

Strikingly, genes within the same TAD tend to have similar gene expression dynamics, suggesting the role of these domains in coordinating the activity of a group of genes (Gibcus and Dekker, 2013). Since their discovery, TADs have been recognized as functional regulatory blocks of genome that prevent erroneous interactions between gene promoters and regulatory elements, restricting contacts to occur between promoters and enhancers within the same TAD (Nora et al., 2012; Hughes et al., 2014; Mifsud et al., 2015; Schoenfelder et al., 2015). Accordingly, deletion of a boundary between two TADs allows for erroneous interactions leading to aberrant gene expression due to non-physiological contacts between genes and enhancers from adjacent TADs (Flavahan et al., 2016; Lupiáñez et al., 2015; Guo et al., 2015; Franke et al., 2016). TADs demonstrate remarkable structural conservation across cell-types and even species (Dixon et al., 2015; Lonfat et al.,

2014; Le Dily et al., 2014; Jin et al., 2013; Dixon et al., 2012) in contrast to A and B compartments, which are largely cell-type specific. Moreover, TADs also display epigenomic uniformity in active or inactive marks (Rao et al., 2014; Sexton et al., 2012; Le Dily et al., 2014). Experimental data revealed that TAD borders are enriched in different features: transcription start sites, different chromatin states and different chromatin-binding proteins, namely architectural protein CTCF and cohesin, hinting to an important role of both proteins in TAD formation and maintenance (Rao et al., 2014; Dixon et al., 2012; Sexton et al., 2012; Nora et al., 2012; Fraser et al., 2015).

Apart from compartments and TADs, a study employing very high resolution Hi-C identified chromatin loops as another sub-chromosomal structure visible as prominent “corner peaks” on Hi-C contact maps (Rao et al., 2014). Loops arise when two loci on the same chromosome are strongly linked together frequently forming a contact domain, an interval of DNA in which all loci pairs between loop anchors interact more frequently with each other than random pairs of loci at similar distances on the linear DNA. The finding of loops gives support to the idea of correlation between chromatin loops and gene regulation by showing that the anchors of cell-type specific loops are often promoters of differentially expressed genes. Interestingly, the loop anchors contain binding sites for CTCF and cohesin. Indeed, many TADs are recognized as loop domains indicating strong contact between CTCF- and cohesin-bound loop anchors (Rao et al., 2017).

The hierarchy of chromatin organization as described in Figure I1 has been recently called into question. Acute depletion of cohesin or CTCF erases chromatin loops and eliminates TADs, while compartments are strengthened (Gassler et al., 2017; Nora et al., 2017; Rao et al., 2017; Wutz et al., 2017; Schwarzer et al., 2017). These results suggest that TADs and compartments are two distinct levels of genome organization established by different mechanisms.

2. Cohesin as a key element for genome organization

As explained in the previous section, CTCF and cohesin are major players in genome organization. CTCF is a DNA binding protein with eleven zinc-finger domains that was identified as a transcriptional regulator of the c-myc oncogene (Lobanenkov et al., 1990). Since then it has been in many functions related to gene regulation (reviewed by Ong and Corces, 2014). CTCF is the main insulator protein described in vertebrates and as such has the ability to interfere with enhancer-promoter communication or establish functional domains of gene expression (Yusufzai et al., 2004). It is highly conserved among eukaryotes

with the notable exceptions of yeast, *D. melanogaster* and *C. elegans*. CTCF binds tens of thousands of sites along mammalian genomes, recognizing a non-palindromic motif (Kim et al., 2007). In this section, I review in detail current knowledge on cohesin.

2.1. Composition and mode of action of cohesin

Cohesin belongs to the *Structural Maintenance of Chromosome* (SMC) family of protein complexes present in eukaryotes. Condensin and the SMC5/6 complex are also members of this family of chromosomal organizers (Jeppsson et al., 2014). Cohesin is composed of SMC1 and SMC3, the kleisin RAD21 and the HEAT repeat containing SA subunit. SMC proteins are long polypeptides (1,000-1,300 amino acids) that consist of two long coiled-coil segments connected by a globular and flexible hinge domain. SMC monomers fold back on themselves at the hinge and establish anti-parallel interactions of the two coiled-coil regions to bring together the N- and C-terminal domains of the protein to create a globular head domain (Figure I4).

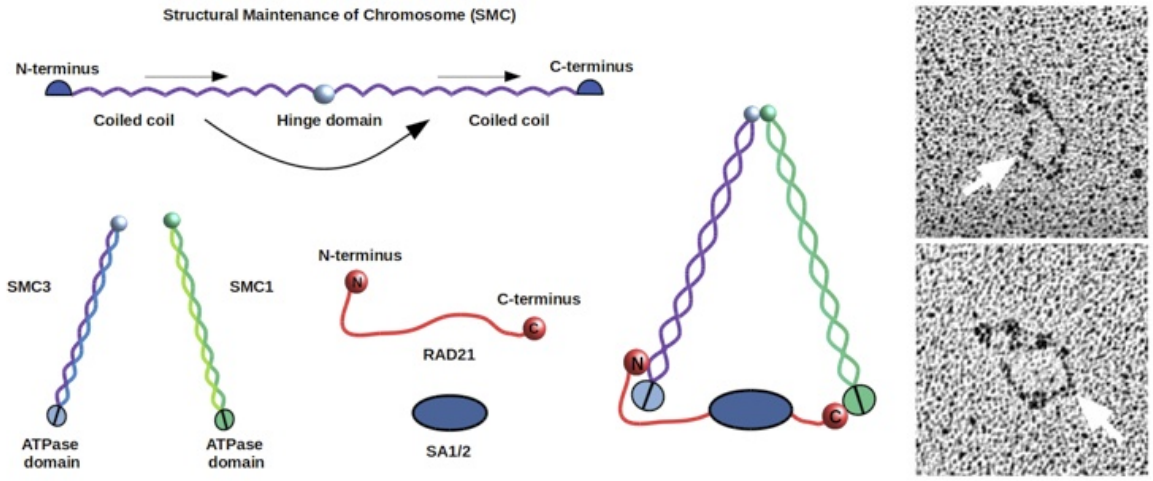


Figure I4. The cohesin complex form a ring

The folding of SMC proteins at the hinge domain brings two termini of the protein together to create the ATPase head domain. Head domains of SMC1 and SMC3 are connected through RAD21, which is in turn bound by the SA subunit. White arrows show electron micrographs of the human cohesin complex purified from HeLa cells. Electron microscopy images taken from Anderson et al., 2002.

Monomers of SMC1 and SMC3 associate tightly with each other through their hinges whereas their head domains are connected by RAD21 and can also engage and interact with each other to form an ABC-like ATPase with two ATP binding pockets. The SA subunit associates to the middle region of RAD21. As we will discuss in detail in Section 4, somatic vertebrate cells express two versions of the SA subunit, SA1 and SA2. Additional versions of

all subunits except SMC3 are also present in germ cells. The complex purified from cell extracts or assembled from recombinant protein has a ring shape, as shown by electron microscopy (Anderson et al., 2002; Huis in 't Veld et al., 2014).

Cohesin subunits were first identified in budding yeast through genetic screens for mutants that led to premature separation of sister chromatids in mitosis (Guacci et al., 1997; Michaelis et al., 1997). Four of these proteins were soon after shown to be part of a holocomplex present in Xenopus egg extracts and human cells that was essential to maintain sister chromatid cohesion (Losada et al., 1998; Losada et al., 2000; Sumara et al., 2000).

It had initially been thought that cohesin might promote cohesion through direct interaction with sister chromatids established through ATPase heads of SMC1 and SMC3 (Nasmyth et al., 2000; Losada and Hirano, 2001). However, the finding that SMC1, SMC3 and RAD21 proteins associate to each other creating a tripartite ring (Haering et al., 2002) together with the observation of ring-shaped complexes in electron micrographs (Anderson et al., 2002) led Kim Nasmyth and colleagues to propose that cohesin might embrace DNA fibers topologically (Gruber et al., 2003). Several biochemical experiments have since supported the idea of cohesin entrapping the two sister chromatids within the lumen of the ring (Haering et al., 2008). One version of this embrace (ring) model holds that a single cohesin ring entraps the two sister chromatids between the SMC coiled coils (Figure I5, left). Another possibility, known as "two gates" model, is that one DNA molecule is held between the coiled coils whereas the second one is kept in the pocket created between the flexible central region of RAD21 and the engaged head domains of SMC1 and SMC3 (Figure I5, middle). A third model suggests that two cohesin rings, each holding a sister chromatid associate through their SA subunits (Zhang et al., 2008) (Figure I5, right).

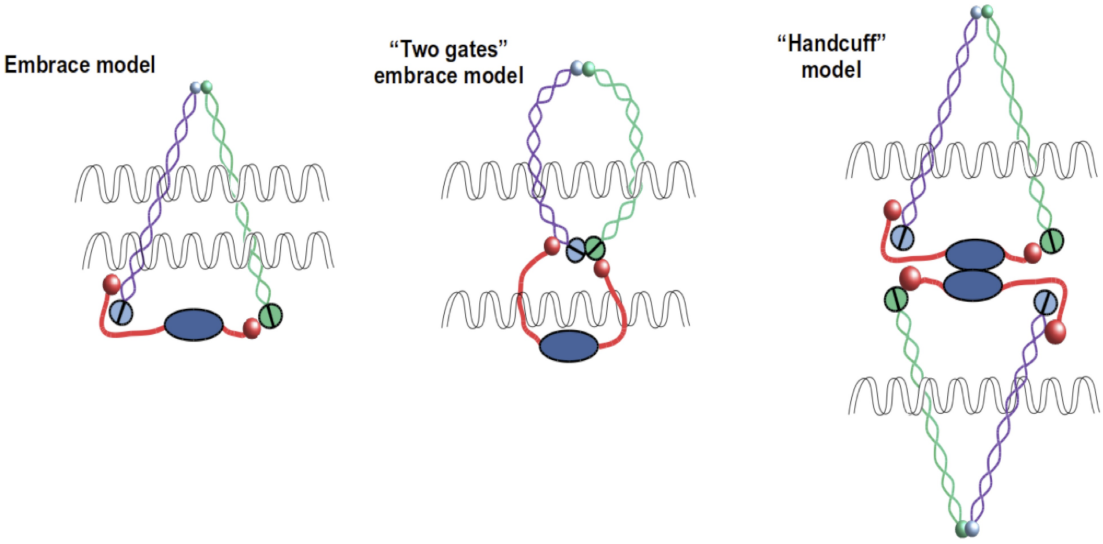


Figure I5. Current models of cohesin association to DNA.

The cohesin ring binds to DNA topologically, holding the chromatin fibers inside its lumen. In the "embrace model", the two sister chromatids are held between the coiled coils whereas the "two gates" model suggests that one sister is held between the coiled coils and the other is located within the pocket created by flexible central region of SCC1 and the engaged SMC head domains. The "handcuff" model assumes proposes that two cohesin complexes each embracing a sister chromatid interact possibly through their SA subunit.

2.2. Association of cohesin with chromatin is dynamic

The association of cohesin with chromatin throughout the cell cycle is tightly regulated and highly dynamic. Cohesin loading occurs early in G1 and is facilitated by the heterodimer of NIPBL and MAU2, which promotes ATP hydrolysis at the SMC heads and causes the dissociation of the hinge domains creating an "entry gate" for the DNA fiber (Arumugan et al 2003; Gruber et al., 2006). In vitro, cohesin can bind DNA topologically even in the absence of the loader, although rather inefficiently (Murayama and Uhlmann, 2014). NIPBL is a hook-shaped protein with several HEAT repeats that is sufficient to stimulate in vitro loading of cohesin on DNA (Chao et al., 2017; Kikuchi et al., 2016) while MAU2 may function as a chromatin adaptor (Chao et al., 2015).

After loading, two other HEAT repeat domain-containing proteins associate with cohesin to modulate its association with chromatin, Pds5 and WAPL. Together, they promote cohesin unloading (Gandhi et al., 2006; Tedeschi et al., 2013; Ouyang et al., 2016) (Figure I6). This process requires dissociation of the interface between the SMC3 head domain and the N terminus of the kleisin subunit, called "exit gate" (Gligoris et al., 2014; Huis In't Veld et al., 2014; Buheitel and Stemmann, 2013).

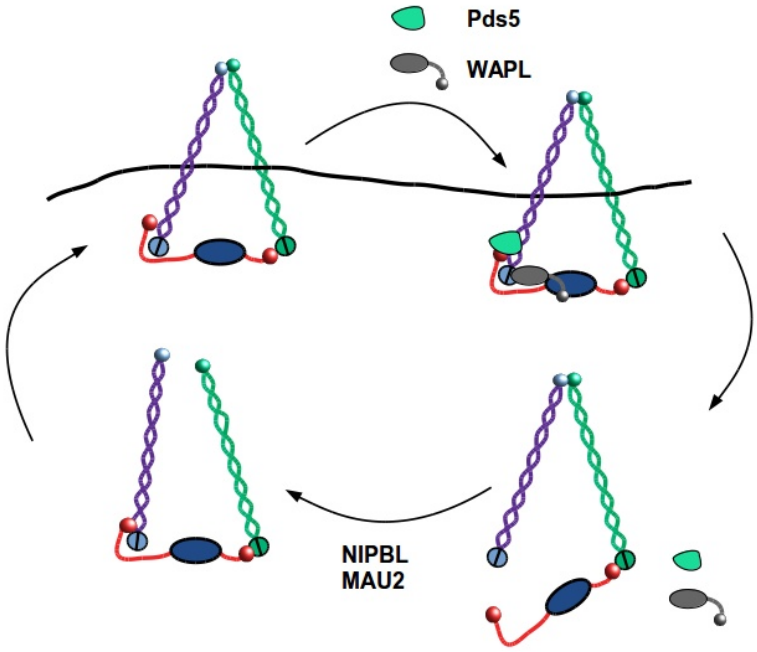


Figure I6. Dynamic association of cohesin to chromatin.

Loading of cohesin is facilitated by the heterodimer NIPBL-MAU2. Association of Pds5-WAPL with chromatin-bound cohesin promotes cohesin unloading.

During DNA replication, a fraction of cohesin, presumably co-entrapping the sister chromatids, becomes stably bound to chromatin (Gerlich et al., 2006). This process requires acetylation of two conserved Lysines (K105 and K108) in the head domain of the SMC3 by acetyltransferases Esco1 and Esco2. Cohesin acetylation is accompanied by binding of Sororin to cohesin through Pds5, which displaces WAPL and counteracts its unloading activity (Nishiyama et al., 2010). When cells enter mitosis cohesin dissociates from chromatin in two steps. More than 90% of cohesin is removed from chromatin in prophase when Sororin dissociation restores WAPL activity and allows opening of the “exit gate” (Losada et al., 2002; Shintomi and Hirano, 2010; Nishiyama et al., 2013). A small fraction of cohesin, enriched around centromeres, is protected from dissociation by a factor called Shugoshin 1 (Sgo1) and its partner the protein phosphatase 2A (PP2A) (Liu et al., 2013). This fraction secures proper chromosome alignment and is removed in anaphase to allow the segregation of the sister chromatids to daughter cells (Waizenegger et al., 2000). In the telophase-early G1 cohesin is loaded again on chromatin, after deacetylation of SMC3 by HDAC8 (Deardorff et al., 2012).

2.3. Cohesion-independent roles of cohesin

As explained above, cohesin was identified and named after its role in sister chromatid cohesion. Cohesion promotes homologous recombination-mediated repair of double-strand breaks during S and G2 phases. In mitosis, cohesion prevents premature separation of sister chromatids until all chromosomes are aligned in the metaphase plate with sister kinetochores attached to opposite spindle poles (Nasmyth and Haering, 2009; Morales and Losada, 2018). Intriguingly, proper cohesion can be maintained with a small fraction of the cohesin complexes present in the cell (Heidinger-Pauli et al., 2010). In addition, cohesin is loaded on chromatin in G1, way before it is required for establishing cohesion, and is present in non-proliferating cells. These facts suggested that cohesin could have other roles in addition to holding the sister chromatids together.

One of the first evidences for cohesion-independent roles of cohesin came from work in Drosophila when mutants for Nipped-B, the ortholog of NIPBL, were found to be deficient in activation of Homeobox genes (Rollins et al., 1999; Rollins et al., 2004). Moreover, cohesin ablation or cleavage in Drosophila postmitotic neurons affected axon pruning through downregulation of the Ecdysone receptor (Pauli et al., 2008; Schuldiner et al., 2008). Cohesin-dependent gene regulation of the Runx1 transcription factor was also reported in zebrafish (Horsfield et al., 2007).

A clear breakthrough in the field came with the analyses of cohesin distribution along human and mouse genomes by chromatin immunoprecipitation, which revealed an extensive colocalization of cohesin with CTCF (Parelho et al., 2008; Wendt et al., 2008). Early work on CTCF recognized this chromatin binding protein as an insulator for its ability to provide insulation of promoters from distant enhancers at the H19-IGF2 locus, a process important for proper expression of these imprinted genes (Bell and Felsenfeld, 2000; Hark et al., 2000). 3C experiments elucidated that CTCF exerts its function by creating allele-specific chromatin loops (Kurukuti et al., 2006; Splinter et al., 2006). Strikingly, it was found that cohesin was indispensable for CTCF to fulfill its enhancer-blocking activity at H19-IGF2 locus (Wendt et al., 2008) as well as in the chicken Beta-globin locus (Parelho et al., 2008). CTCF was dispensable for cohesin loading to chromatin, but it was required for its recruitment to specific loci (Wendt et al., 2008).

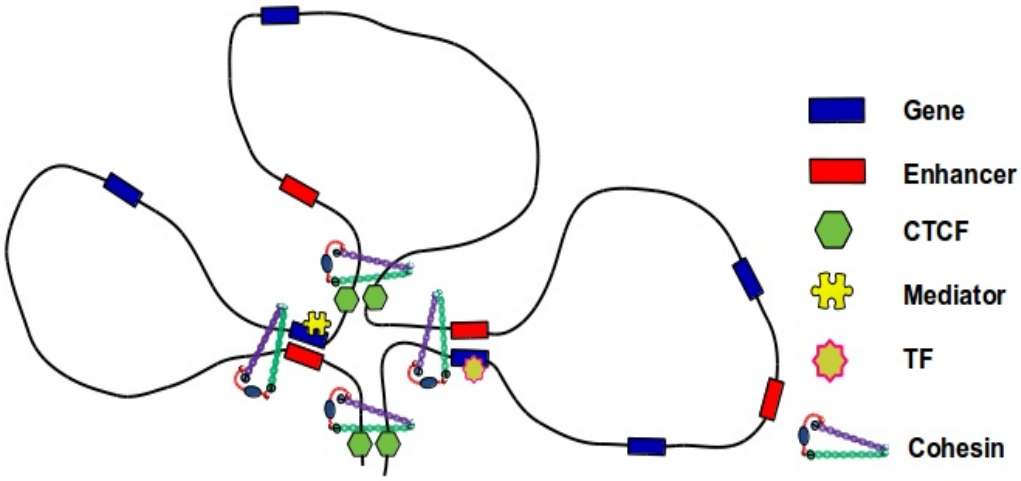


Figure I7. Cohesin organizes interphase chromatin through looping

Cohesin can tether two distant genomic loci in-cis promoting long-range genomic interactions. Together with CTCF, Mediator complex and other TFs, cohesin mediates loops between genes and regulatory elements. Moreover, cohesin-mediated loops can regulate transcription of gene clusters and organize replication factories (not depicted).

The ability of cohesin to bring together two DNA molecules in *trans*, the sister chromatids, sparked the idea that cohesin could also generate chromatin loops by bringing together distant genomic loci of one chromatid, i.e. *in cis,* that could control enhancer–promoter interactions topologically, maybe together with CTCF (Figure I7). Evidence supporting this hypothesis came from 3C studies showing that cohesin mediates long-range chromosomal cis-interactions at the developmentally regulated cytokine locus IFNG (Hadjur et al., 2009). Cohesin and CTCF were also reported to regulate gene expression through chromatin looping in gene clusters such as the human apolipoprotein gene cluster (Mishiro et al., 2009) and protocadherin-alpha gene cluster (Monahan et al., 2012). Additional results using HiC and 4C analyses showed the importance of cohesin-CTCF for chromatin contacts and overall chromosome organization (Sofueva et al., 2013; Zuin et al., 2014)

Cohesin can also function in genome organization and gene regulation in a CTCF-independent manner. Analysis of cohesin distribution in liver and breast cancer cells revealed a fraction of cohesin positions lacking CTCF that were instead bound by tissue specific transcription factors (Schmidt et al., 2010). Likewise, cohesin was shown to occupy the promoters of pluripotency genes in mouse embryonic stem cells together with Mediator and NIPBL and to mediate contacts with the corresponding enhancers to promote high gene expression levels (Kagey et al., 2010).

Genome topology is also important for processes other than transcriptional regulation. In preparation for DNA replication, neighboring origins of replications cluster together to form

replication factories through chromatin looping (Jackson and Pombo, 1998; Lemaitre et al., 2005). Cohesin was found to interact with the components of pre-RC (pre-replication complex) to spatially organize these replication factories enabling efficient origin firing (Guillou et al., 2010). Cohesin also participates in DNA looping required for locus rearrangement at the Immunoglobulin and T cell receptor loci in B and T lymphocytes, respectively (Degner et al., 2011; Seitan et al., 2011).

2.4. Genome-wide distribution of cohesin: implications for genome organization

The first published genome wide map of cohesin distribution in a metazoan organism came from Drosophila (Misulovin et al., 2008). It showed a complete colocalization of cohesin with its loader (Nipped-B), peaking at transcription start sites of active genes. In mammals, however, only a small fraction of cohesin positions was found at the same sites as the loader (Zuin et al., 2014) whereas, as explained above, the overlap with CTCF was very extensive (Wendt et al., 2008; Parelho et al., 2008). From these data, it was not clear whether the overlap of cohesin with CTCF was the consequence of cohesin being loaded at CTCF-bound locations or the ability of cohesin to move laterally after binding DNA topologically as it had been previously suggested from experiments in yeast (Hu et al., 2011; Ocampo-Hafalla and Uhlmann, 2011).

Consistent with the latter possibility, lateral movement of cohesin complexes along DNA by passive diffusion has been recently observed in vitro using single molecule imaging (Davidson et al., 2016; Stigler et al., 2016). DNA-bound cohesin acquires a conformation that reduces the size of the central pore of the complex thus hindering its ability to diffuse freely past different DNA-bound obstacles, such as CTCF. How well these assays reflect cohesin dynamics in vivo is uncertain as they are performed on naked DNA and do not require ATP or the cohesin loader. A third study showed cohesin complexes loaded by NIPBL-MAU2 and moving along DNA in an ATP-dependent manner (Kanke et al., 2016). Addition of WAPL-Pds5 reduced translocation while acetylation promoted it. The underlying reason for these effects remains unclear, although it could be related with conformational changes in the ring affecting its ATPase activity.

Movement of cohesin in vivo is inferred from its distribution, assessed by ChIP-seq. In mouse fibroblasts lacking CTCF, cohesin accumulation at CTCF sites was reduced and the complex appeared instead at the transcription start sites of active genes, where NIPBL was also found and where loading possibly occurred. Co-depletion of CTCF and WAPL to prevent cohesin unloading led to cohesin accumulation in broad peaks at 3′-ends of active

genes, in particular if these converge on each other, but not when transcription was switched off (Busslinger et al., 2017). These results suggest that transcriptional machinery can push cohesin in mammalian cells, as previously reported in yeast (Lengronne et al., 2004; Glynn et al., 2004). In vitro, motor proteins can also push cohesin along DNA (Stigler et al., 2016). Some evidences also suggest that cohesin can be pushed by the replication machinery (Kanke et al., 2016).

2.5. The loop extrusion model

In addition to this passive movement, current evidence support an active translocation of cohesin along DNA or, more exactly, active reeling of DNA through cohesin. It was hypothesized several years ago that SMC complex could act by catching a small loop that would then be processively enlarged by the sliding of the complex down the base of the loop (Nasmyth, 2001).

Chromatin loops predominantly form between convergent CTCF sites, and such orientation is required for looping (de Wit et al., 2015; Guo et al., 2015; Sanborn et al., 2015), suggesting a linear scanning mechanism that recognizes the orientation of CTCF sites. Based on polymer simulations, two groups proposed that genome organization as observed by Hi-C could be modeled with *cis*-acting loop-extruding factors, likely cohesin, forming progressively larger loops until being stopped by boundary proteins, like CTCF (Sanborn et al., 2015; Fudenberg et al., 2016). Very recent work has now provided experimental support for this loop-extrusion model (Figure I8). For instance, a prediction of the model is that the longer the extruding complex stays bound to DNA, the longer the loops should be. Indeed, depletion of WAPL or Pds5 to prevent cohesin unloading increases loop length genome wide (Haarhuis et al., 2017; Wutz et al., 2017). How cohesin recognizes the directionality of CTCF is currently unknown. It is also unclear if a single complex or two tethered complexes are responsible for the extruding. Importantly, in recent in vitro experiments using purified condensin showed that a single complex performs loop extrusion: it reeled one side of the DNA loop while the other end remained stably bound (Ganji et al., 2018).

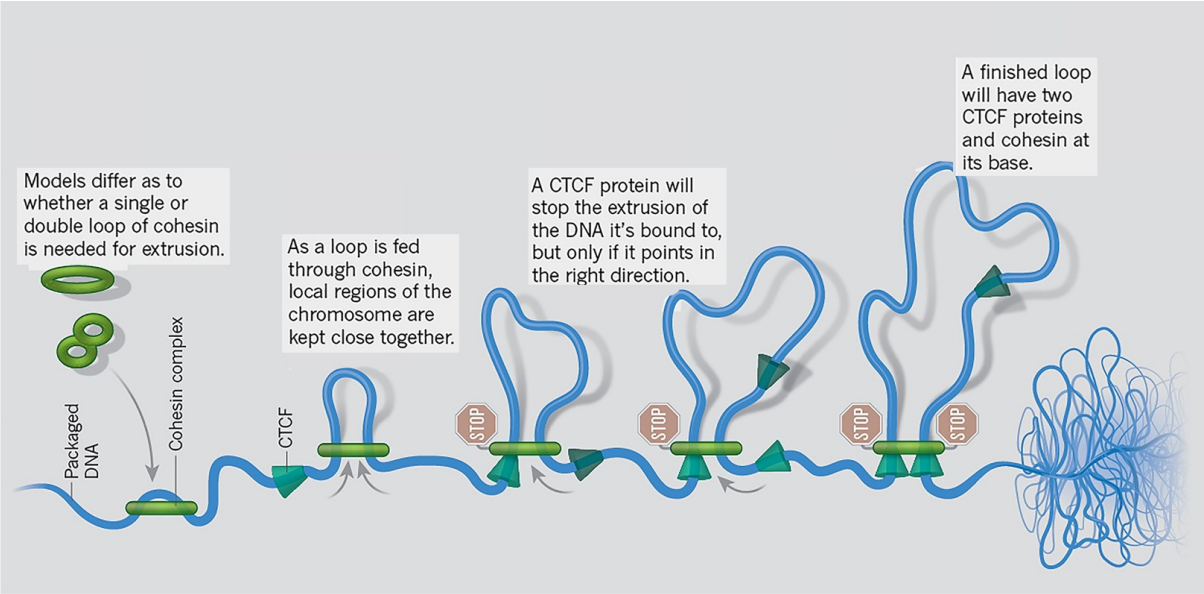


Figure I8. Cohesin as a loop extrusion factor

The loop extrusion model for genome organization helps to explain how certain regions of chromosomes stay close together and why cohesin rings are often found coinciding with CTCF, which binds to specific DNA sequences. Adapted from Dolgin, 2017.

3. Cohesin mutations in human disease

The fact that cohesin plays major roles in chromosome segregation and genome organization makes it rather difficult to understand the pathological consequences of cohesin defects. In the germline, mutations in cohesin genes and its regulators are responsible for developmental abnormalities collectively called cohesinopathies. Mutations in cohesin have also been identified in many different cancer types.

3.1. Cohesin mutations in cohesinopathies

Most common cohesinopathies are Cornelia de Lange syndrome (CdLS) and Roberts Syndrome (RBS), a multisystemic disorders with a broad clinical spectrum that includes cognitive impairment, growth retardation, craniofacial abnormalities, limb defects, deafness and heart defects among others (Banerji et al., 2017a).

CdLS affects up to 1:10,000 newborns. This autosomal dominant disorder is caused in 65% of the cases by mutations in NIPBL gene, while mutations in SMC1, SMC3, RAD21, and HDAC8 account for around 10% of diagnosed patients. Cells from these patients do not display obvious cohesion defects, and instead altered transcriptional profiles are observed (Liu et al., 2009). Studies using zebrafish and mouse models revealed several pathways

crucial for embryonic development, such as the *HOX* genes that might be affected by cohesin insufficiency while defects in cohesion were not found (Remeseiro et al., 2013). In NIPBL heterozygous mice the protocadherin genes (*PCDH*), clusters of genes that are expressed in an exclusive manner to provide neuronal identity, were found to be misregulated (Kawauchi et al., 2009) and subsequent studies showed that cohesin-mediated DNA looping is involved in *PCDH* promoter choice (Guo et al., 2012; Cuadrado et al., 2015) indicating how cohesin insufficiencies might affect the developing brain.

Recently, a new genetic disorder was discovered, the CHOPS syndrome (C for cognitive impairment and coarse facies, H for heart defects, O for obesity, P for pulmonary involvement and S for short stature and skeletal dysplasia) that shares striking similarities with CdLS in terms of transcriptional profiles (Izumi et al., 2015). It arises due to gain-of function mutation in the gene coding for AFF4, a subunit of super elongation complex (SEC) responsible for mobilization of paused RNA polymerase II (RNA pol II). These mutations led to aberrant genome-wide distribution of both AFF4 and cohesin around TSS of upregulated genes. Interestingly, similar alterations in the behavior of AFF4 and cohesin were observed in CdLS, pointing to a similar mechanism of pathogenesis in both syndromes related with the role of cohesin in transcriptional regulation.

RBS is caused by homozygous mutations in the gene for Esco2, one of the two cohesin acetyltransferases present in vertebrate cells (Vega et al., 2005). Unlike CdLS, cells from RBS patients exhibit a spectrum of cohesion aberrations: chromosomal segregation defects, premature centromere separation, aneuploidy, lagging chromosomes and reduced proliferation. Esco2 appears to be particularly important for cohesion establishment in pericentric heterochromatin (Whelan et al., 2012). Therefore, the proposed model for RBS holds that the phenotype is due to mitotic failure and loss of progenitor stem cells through apoptosis.

3.2. Cohesin mutations in human malignancies

Cohesin had first been associated to cancer when mutations in genes encoding cohesin subunits were linked to genome instability observed in colorectal cancer (Barber et al., 2008). A few years later, mutations in *STAG2* gene were found in glioblastoma, Ewing sarcoma and melanoma (Solomon et al., 2011). More importantly, this study also hinted to chromosome mis-segregation as the consequence of cohesin dysfunction. Recently, pan-cancer studies identified recurrent somatic mutations in the genes encoding cohesin subunits and cohesin regulator proteins across many cancer types, recognizing cohesin

network as one of the most frequently mutated in cancer (Kandoth et al., 2013; Lawrence et al., 2013; Leiserson et al., 2015; Martincorena et al., 2017) (Figure I9). Contrary to the initial studies mentioned above, more recent data indicate that in the majority of tumours or cancer cell lines analyzed there is no clear correlation between the presence of cohesin mutations and aneuploidy. Thus, cohesion is unlikely to be the function impaired in these tumour cells. Instead, lower cohesin dosage could affect differentiation or alter gene expression.

Sequencing efforts in acute myeloid leukemia (AML) samples and other myeloid malignancies identified mutations in cohesin related genes in more than 10% of the tumours (Welch et al., 2012; Yoshida et al., 2011; The Cancer Genome Atlas Research Network, 2013). These mutations were found to be largely nonsense (70%) or frame-shift (17%) with no obvious hotspots, suggesting the molecular loss of function as the probable consequence (Kon et al., 2013; Thol et al., 2014). Even in the absence of cohesin mutations, low expression of cohesin components was detected in a significant fraction of myeloid malignancies (Thota et al., 2014). Clonal analysis showed that cohesin mutations are early events and suggested that mutations are present in the dominant clone, but not in the founder one (Jan et al., 2012; Thol et al., 2014). These results indicate that cohesin mutations provide a specific genetic context for the acquisition of the “second hit” or promote transformation to more aggressive disease.

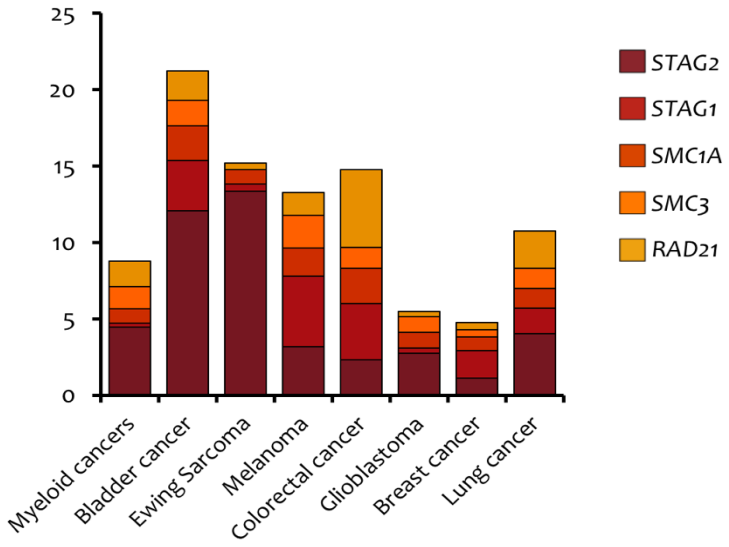


Figure I9. Cohesin mutations in cancer

Data were obtained from the cBioPortal for Cancer Genomics. Only mutations in the genes encoding cohesin subunits are shown. Additional mutations are found in cohesin regulators. Adapted from De Koninck and Losada, 2016.

Importantly, cohesin mutations in myeloid malignancies were not associated with aberrant karyotypes (Thota et al., 2014). To better understand the contribution of cohesin mutations to tumorigenesis, functional studies were performed using transgenic mouse models and shRNA knockdown (Mazumdar et al., 2015; Mullenders et al., 2015; Viny et al., 2015). These studies consistently demonstrated that cohesin mutation or loss led to in-vivo and in-vitro expansion of hematopoietic stem cells (HSC), impaired differentiation and skewing of transcriptional program towards HSC signature. Moreover, changes in chromatin accessibility were also observed in regions enriched in DNA-binding motifs of transcription factors involved in the maintenance of the HSC stem-cell programme (Mazumdar et al., 2015; Mullenders et al., 2015; Viny et al., 2015; Wilson et al., 2010; Fisher et al., 2017). Importantly, even when cohesin dosage was drastically reduced, sister chromatid cohesion seemed to be fine, as previously reported in yeast (Heidinger-Pauli et al., 2010). Thus, these studies suggested that reduced cohesin activity promotes transformation through delaying lineage commitment while favoring activation of stem cell transcriptional programs, and that this could be achieved through alteration of chromatin accessibility of stemness associated TFs. However, it is unclear whether the observed changes in gene expression and chromatin accessibility are cause or consequence of the observed increase in self-renewal and the reduced differentiation.

Cohesin mutations are also common in urothelial bladder cancer. While other cohesin subunits were found mutated at low rates (2-6%), the mutation rate of *STAG2,* the gene encoding SA2*,* was substantially higher (16%) (Balbás-Martínez et al., 2013; Guo et al., 2013; Solomon et al., 2013; Taylor et al., 2014). *STAG2* loss in low-grade, non-aggressive tumors was associated to better prognosis, whereas the opposite was observed for more aggressive tumors. Again, no clear link between *STAG2* loss and aberrant karyotypes could be found. The same is true for Ewing sarcoma, a pediatric tumor of the bone and soft tissues in which *STAG2* mutations are also very frequent (Brohl et al. 2014; Crompton et al. 2014; Tirode et al. 2014).

Finally, whole-genome sequencing data from more than 200 samples of colorectal cancer (CRC) patients together with ChIP-seq analyses in a CRC cell line revealed a high incidence of mutations in cohesin/CTCF binding sites in the non-coding genome (Katainen et al., 2015). A fraction of these mutations are predicted to affect CTCF binding affinity to cis-regulatory elements and could therefore contribute to tumourigenesis through aberrant expression of their target genes. Another epigenetic mechanism recently described in

gliomas involves disruption of boundary elements through hypermethylation of CTCF/cohesin binding sites leading to oncogene activation (Flavahan et al., 2016).

4. Two versions of cohesin in somatic cells with non redundant functions

Somatic vertebrate cells contain two versions of the cohesin complex that carry SMC1, SMC3, RAD21 and either SA1 or SA2 (Losada et al., 2000; Sumara et al., 2000). Initially, cohesin-SA2 was shown to be at least 3 times more abundant than SA1 in human and Xenopus somatic cells, while this ratio was reversed in extracts prepared from Xenopus oocytes. Thus, the relative abundance of the two variant complexes could be regulated during development or according to cell types (Losada et al., 2000). The functional specificity of the two variant cohesin complexes remains unclear and is one of the main goals of the current Thesis.

Both SA1 and SA2 proteins are more than 1250 amino acids long and are characterized by the presence of 17 HEAT (Huntingtin, elongation factor 3 (EF3), protein phosphatase 2A (PP2A), and the yeast kinase TOR1) repeats in the central region (Hara et al., 2014). They share high sequence homology along the central region, but differ in their C and N terminal regions (Figure I10). Current evidence suggests that the interaction between cohesin and CTCF overlaps largely with SA motif and is found to be nearly perfectly conserved between two paralogs (Xiao et al., 2011)

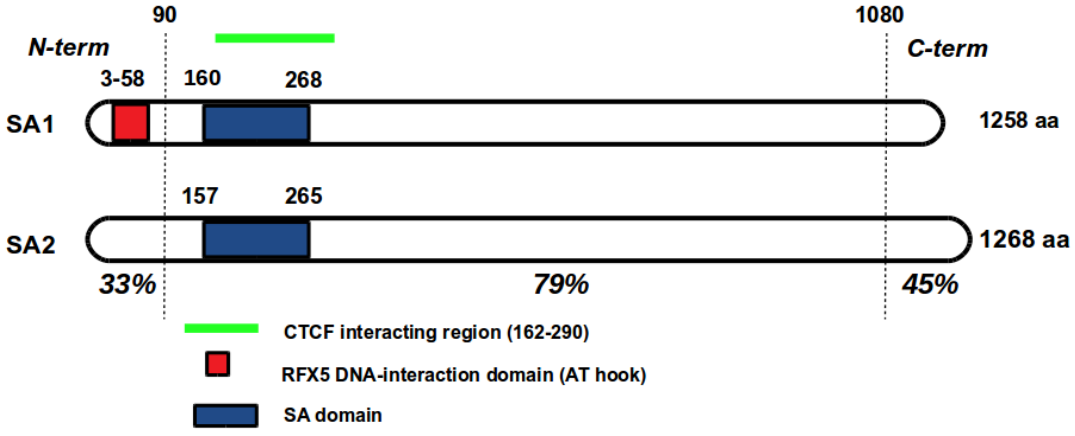


Figure I10. Schematic representation of SA proteins and their domains

The percentages refer to sequence identity between two proteins in the corresponding regions (N-terminal, central and C-terminal regions). Modified from Pezic et al., 2017).

The crystal structure of SA2 bound to the central flexible regions of RAD21 was recently reported (Hara et al., 2014). It revealed an extensive interface between SA2 and RAD21, which shows high conservation between SA1 and SA2. The study further identified a region

from aa 250 to aa 330 to which cohesin regulators WAPL and Sgo1 bind in a mutually exclusive manner, providing a molecular mechanisms for centromeric cohesin protection by Sgo1 in mitosis. However, the crystal structure does not contain the N and C terminal regions that are less conserved between the two SA proteins.

In spite of their high sequence similarity and their similar chromatin-association dynamics throughout the cell-cycle, cohesin-SA1 and cohesin-SA2 have non-redundant functions. The best evidence for this assertion is that mouse deficient for SA1 or SA2 are embryonic lethal (Remeseiro et al., 2012a; M. De Koninck and E. Lapi, unpublished results). Analysis of SA1 deficient mouse embryo fibroblasts (MEFs) revealed a more pronounced role of cohesin-SA1 in telomeric cohesion while cohesin-SA2 appeared to be more important for cohesion at centromeres. Depletion of SA1 and SA2 using siRNA in HeLa cells had reached similar conclusions (Canudas and Smith, 2009). SA1 contains an AT-hook domain, absent in SA2, shown to be essential for the association of SA1 with telomeric DNA (Bisht et al., 2013; red square in Figure I10) and interacts with the telomere binding protein TRF1 (Canudas et al., 2007). Telomere cohesion is relevant to allow efficient replication of telomeres and in the absence of SA1 replication defects lead to chromosome segregation defects and increased aneuploidy compared with wild type MEFs (Remeseiro et al., 2012a). Despite the preferential roles of cohesin-SA1 and cohesin-SA2 in mediating cohesion at different chromosome regions, any of the two complexes provide sufficient cohesion to allow cell proliferation, at least in cultured cells (van der Lelij et al., 2017).

SA1 deficient murine embryos died by 12.5 dpc although some exceptional embryos reached later stages of development. These embryos displayed severe developmental abnormalities at the organismal level as well as gene expression changes at the cellular level (Remeseiro et al., 2012b). Moreover, mapping cohesin distribution in wild type and SA1 KO MEFs revealed that cohesin-SA2 repositions to new sites with reduced overlap with CTCF in the absence of cohesin-SA1. We speculate that these changes in cohesin distribution likely contribute to alter gene expression.

Co-immunoprecipitation of cohesin-SA1 but not cohesin-SA2 with RNA polymerase II (RNA pol II) and components of the superelongation complex (SEC) from HeLa cell lysates has been reported (Izumi et al., 2014). This complex is responsible for mobilization of the paused RNA pol II machinery and thereby regulates transcriptional elongation. Results in Drosophila also suggested a role for cohesin in promoting the transition promoter-proximal paused RNA polII to elongation at many genes although in this organism a single SA subunit exists (Schaaf et al., 2013). As mentioned before, mutations in components of the SEC complex

have been identified in CHOPS syndrome, a disorder that phenotypically overlaps with Cornelia de Lange syndrome (CdLS). It is therefore possible that some of the transcriptional alterations observed in SA1 deficient cells result from altered regulation of transcriptional elongation.

Reduced dosage of SA1 in adult animals results in increased susceptibility to spontaneous cancer, in particular pancreatic cancer (Remeseiro et al., 2012a). Interestingly, reduced dosage of SA1 altered the architecture of the Regenerating islet-derived (Reg) gene cluster in the pancreas of adult mice, as probed by 4C-seq analyses, and decreased the expression of Reg genes (Cuadrado et al., 2015). Given the role of Reg proteins in inflammation, and the known impact of inflammation in pancreatic cancer development, we proposed that downregulation of Reg gene expression could be responsible for the increased incidence of pancreatic cancer observed SA1 heterozygous mice.

As reviewed in the previous section, loss of function mutations in the *STAG2* gene have been identified in several human cancers most prominently bladder cancer, Ewing sarcoma and acute myeloid leukaemia (De Koninck and Losada, 2016). Genomic instability was not observed in these tumours, which suggested that loss of cohesin-SA2 likely contributes to tumourigenesis through altered gene regulation. Functional studies in hematopoietic stem cell precursors support this idea (Viny et al., 2015; Viny et al., 2018). Complete elimination of cohesin is lethal for the cell while the presence of either cohesin variant is sufficient to allow proliferation of tissue culture cells (van der Lelij et al., 2017), but not during embryo development. We speculate that cohesin-SA1 can carry out the essential functions of cohesin related with sister chromatid cohesion in *STAG2*-deficient cancer cells while it may not be able to fully compensate for the loss of cohesin-SA2, thereby promoting tumourigenesis through gene deregulation.

Little is known about the potential functional specificities of cohesin-SA1 and cohesin-SA2 in genome organization and gene regulation. Addressing this question is the main aim of this Thesis.

Objectives

1. **Analyse the genome-wide distribution of cohesin-SA1 and cohesin-SA2 as well as their association to functional genomic elements in non tumoral human cell lines.**

2. **Evaluate the stability and dynamics of association to chromatin of cohesin-SA1 and cohesin-SA2.**

3. **Assess the impact of cohesin-SA1 and cohesin-SA2 downregulation on the transcriptome.**

4. **Study the specific contributions of cohesin-SA1 and cohesin-SA2 to 3D genome organization.**

Materials and Methods

Cell lines. Human primary cell lines were purchased from Lonza and cultured according to the manufacturer's recommendations. NHA (Normal Human Astrocytes, CC-2565) were grown in ABM basal medium (CC-3187) supplemented with AGM Bulletkit (CC-4123); SKMC (Skeletal Muscle Cells, CC-2561) were cultured in SkBM basal medium (CC-3161) supplemented with SkGM Bulletkit (CC-4139); NHBE (Normal Human Bronchial Epithelial Cells, CC-2540) were cultured in BEBM basal medium (CC-3171) supplemented with BEGM Bulletkit (CC-4175); HCAEC (Coronary Artery Endothelial Cells, CC-2585) were grown in EBM2 basal medium (CC-3156) supplemented EGM2-MV Bulletkit (CC-4147); NHEK (Normal Human Epidermal Keratinocytes, #00192627) were grown in KBM-Gold basal medium (#00192151) supplemented with KGM-Gold Bulletkit (#00192060). HMEC (Normal Mammary Epithelial Cells, CC-2551) were cultured in MEBM basal medium (CC-3171) supplemented with MEGM Bulletkit (CC-3150). NHOst (Normal Human Osteoblasts, CC-2538) were grown in OBM basal medium (CC-3208) supplemented with OGM Bulletkit (CC-3207). PrEC (Prostate Epithelial Cells, CC-2555) were cultured with PrEBM basal medium (CC-3165) supplemented with PrEGM Bulletkit (CC-3166). HUVEC (Human Umbilical Vein Endothelial Cells, CC-2517) were grown in EBM basal medium (CC-3121) supplemented with EGM Bulletkit (CC-3124). MCF10A cells (a gift from M. Quintela, CNIO) were grown in DMEM/F12 (#31330038, Thermofisher) supplemented with 20ng/ml of EGF, 0.5mg/ml hydrocortisone, 100ng/ml of cholera toxin, 10mg/ml of insulin and 5% of horse serum.

Antibodies. A rabbit polyclonal antibody recognizing human WAPL was generated using a recombinant C-terminal fragment of the protein (352 amino acids long) obtained by PCR amplification of full length hWAPL cDNA [a gift from T. Hirano (RIKEN, Japan)]. Rabbit polyclonal antibodies against SA1, SA2, SMC1 and RAD21 were generated using synthetic peptides as immunogens [SA1-C (CEDDSGFGMPMF), SA2 (CDPASIMDESVLGVSMF), SMC1 (CDLTKYPDANPNPNEQ), RAD21, GGDQDQEERRWNKRTQQMLC] and affinity purified. Rabbit polyclonal antibody against ZMYM2 was generated using a ZMYM2 fragment (residues 923–1377) as immunogen [a gift from H. Yu (UT Southwestern, US)]. Commercial antibodies were CTCF (07-729, Millipore), tubulin (DM1A, Sigma), histone H3 (Abcam AB1791).

Quantitative immunoblotting in whole cell extracts and chromatin fractions. Cells were collected by trypsinization, counted, resuspended in SDS-PAGE loading buffer at 10^7 cells/ml, sonicated and boiled. Equal volumes were separated by SDS-PAGE and analyzed by immunoblotting. For chromatin fractionation we followed the protocol from (Méndez and Stillman, 2000). To isolate chromatin, cells were resuspended (4×10^7 cells/ml) in buffer A

(10 mM HEPES, [pH 7.9], 10 mM KCl, 1.5 mM $MgCl_2$, 0.34 M sucrose, 10% glycerol, 1 mM DTT, 5 µg of aprotinin per ml, 5 µg of leupeptin per ml, 0.5 µg of pepstatin A per ml 0.1 mM phenylmethylsulfonyl fluoride). Triton X-100 (0.1%) was added, and the cells were incubated for 5 min on ice. Nuclei were collected in pellet 1 (P1) by low-speed centrifugation (4 min, 1,300 × *g*, 4°C). Nuclei were washed once in buffer A, and then lysed in buffer B (3 mM EDTA, 0.2 mM EGTA, 1 mM DTT, protease inhibitors as described above) 30 min on ice. Insoluble chromatin was collected by centrifugation (4 min, 1,700 × *g*, 4°C), washed once in buffer B, and centrifuged again under the same conditions. Chromatin pellet was resuspended in Laemmli buffer and sonicated for 15 s at 20% amplitude. Fractions were run on SDS gels alongside increasing amounts of recombinant proteins corresponding to C-terminal fragments of human SA1 and SA2 to estimate the amount of each variant subunit. To assess the strength of chromatin association of cohesin variants, chromatin fractions were treated with modified buffer A (10mM HEPES, 1.5mM MgCl2, 0.34M sucrose, 10% glycerol, 1mM DTT and protease inhibitors) containing 0.25 or 0.5M NaCl for 10, 20 or 30 min on ice. Solubilized proteins were separated from insoluble chromatin by low speed centrifugation (4 min at 1,700 x g) and the latter analyzed by immunoblotting.

Immunoprecipitation. Whole cell extracts from MCF10A cells were prepared by lysis on ice for 30 min in TBS supplemented with 0.5% NP-40, 0.5mM DTT, 0.1mM PMSF and 1X complete protease inhibitor cocktail (Roche) followed by sonication. NaCl was added to 0.3M and the extract rotated for 30 min at 4ºC. After centrifugation, the soluble fraction was recovered and diluted to bring the extract back to 0.1M NaCl and 10% glycerol was added. Antibodies were cross-linked to protein A Pureproteome magnetic beads (Millipore) at 1 mg/ml (SA1, SA2 and IgG as control) and incubated with extracts overnight at 4ºC. The beads were washed 6 times with 20 vol of lysis buffer and anlaysed by LC-MS/MS.

LC-MS/MS Analysis. Proteins from antibody beads were eluted in two consecutive steps in 2 vol of elution buffer (8M urea, 100mM Tris-HCl pH 8) by shaking for 10 min. Samples were digested by standard Filter Aided Sample Preparation (FASP). Proteins were reduced with 10mM DTT, alkylated with 50mM IAA for 20 min in the dark and digested with 1:50 Lys-C (Wako) for 4 h. Samples were diluted in 50mM ammonium bicarbonate and digested with 1:100 Trypsin (Promega) overnight at 37ºC. Resulting peptides were desalted using a Sep-Pak C18 cartridge for SPE (Waters Corp.), vacuum-dried and resuspended in 0.5% FA. Immunoprecipitates were analysed using a nanoLC Ultra system (Eksigent, Dublin, CA) coupled with a LTQ-Orbitrap Velos instrument (Thermo) via nanoESI (ProxeonBiosystem, Waltham, MA). Two technical replicates were performed. Raw data were analysed using MaxQuant1.5.3.30 with Andromeda as the search engine against UniProtKB/Swiss-Prot

(20,584 sequences). Peptides were filtered at 1% FDR. For protein assessment (FDR <1%), at least one unique peptide was required for both identification and quantification. Other parameters were set as default. The resulting “proteingroup.txt” file was loaded in Perseus (v1.5.1.6). Missing values were imputed from a normal distribution. A two-sample Student’s T-Test (one side) was used corrected for multiple testing using a permutation-based approach.

siRNA. MCF10A cells were transfected with 50 nM onTARGETplus SMARTpool siRNAs (Dharmacon L-010638, L-021351, L-006833 and L-020165 for SA1, SA2, SMC1 and CTCF, respectively) using DharmaFECT reagent 1. Transfection efficiency was first estimated by qRT-PCR 24 h after transfection, and typically reached more than 90% downregulation (data not shown). Cells were taken at 72h and protein levels assessed by immunoblot.

ChIP sequencing and analysis. Chromatin immunoprecipitation (ChIP) was performed as described (D’Alessio et al., 2015), with some modifications. Confluent cells were cross-linked with 1% formaldehyde added to the media for 15 minutes at RT. After quenching with 0.125M Glycine, fixed cells were washed twice with PBS containing 1mM PMSF and protease inhibitors, pelleted and lysed in lysis buffer (1%SDS, 10mM EDTA, 50mM Tris-HCl pH 8.1) at $2x10^7$ cells/ml. 10^7 cells equivalent to 40-50 mg of chromatin were used per immunoprecipitation reaction with 25 mg of antibody. Sonication was performed with a Covaris system (shearing time 30 min, 20% duty cycle, intensity 6, 200 cycles per burst and 30 s per cycle) in a minimum volume of 2 ml. For calibrated ChIP-seq in siC, siSA1 and siSA2-treated MCF10A cells, 20% of chromatin from mouse ES cells was added to the human chromatin. The calibrated ChiP-seq was performed in order to correct the technical variability between replicates and obtain quantitative information about occupancy of the same antibody among samples from different conditions (Figure MM1). We doubled the amount of antibody used for the immunoprecipitations in order to reduce differences on antibody saturation among conditions. ChIP-seq profiles for each antibody were multiplied by the occupancy ratio (OR) = $(W_m IP_h)/(W_h IP_m)$, where W_h and IP_h are the number of reads mapped to the human genome from input (W) and immunoprecipitated fractions (IP) and W_m and IP_m are reads mapped to the mouse genome from input and IP fractions (Hu et al., 2015).

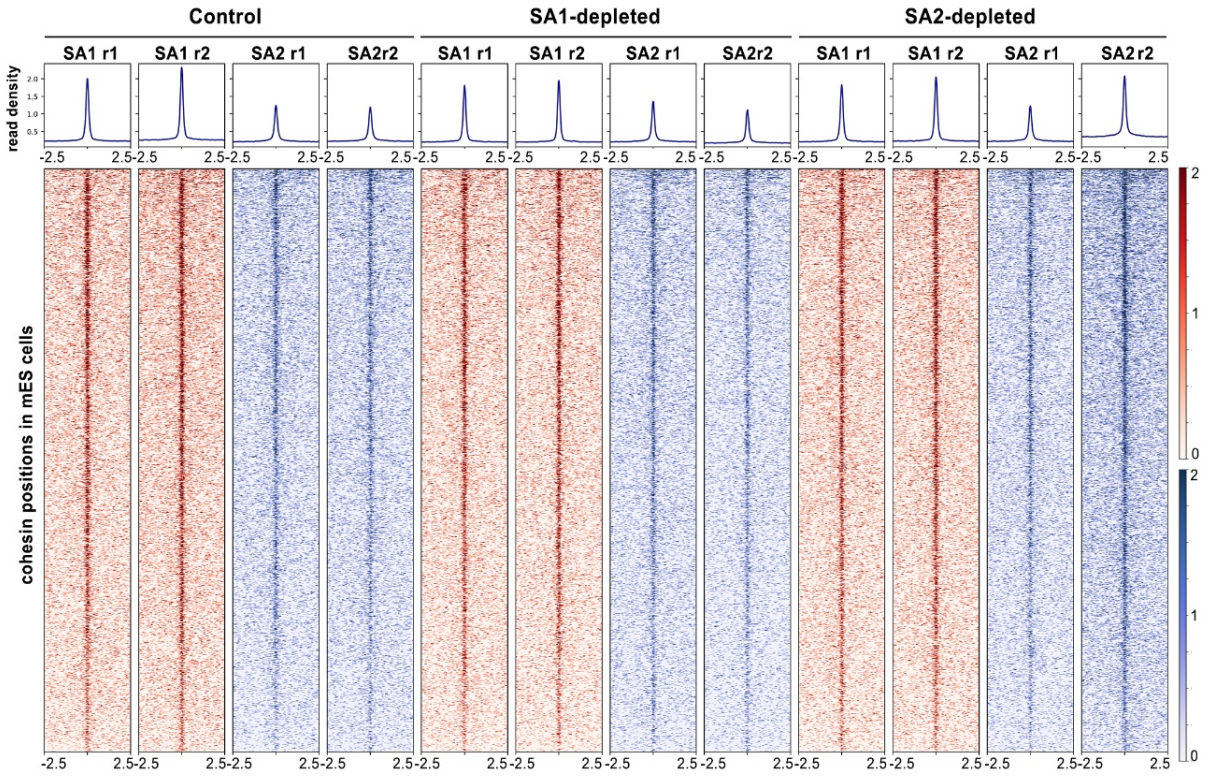


Figure MM1. Calibrated ChIP-seq read distribution around common positions defined in mouse ES cells

Calibrated ChIP-seq allows for quantitative comparison of protein binding between samples of different conditions. It uses “calibration” sample as an internal control to correct for technical noise. Note the experimental variability of the two replicates for SA2 in SA2-depleted cells.

From 6 to 10ng of immunoprecipitated chromatin (as quantitated by fluorometry) were electrophoresed on an agarose gel and independent sample-specific fractions of 100–200 bp were taken. Adapter-ligated library was completed by limited-cycle PCR with Illumina PE primers (11 to 13 cycles). DNA libraries were applied to an Illumina flow cell for cluster generation and sequenced on the Illumina Genome Analyzer IIx (GAIIx). Image analysis was performed with Illumina Real Time Analysis software (RTA1.8).

Alignment of 50-bp (76-bp for calibrated ChIP samples) long sequences to the reference genome (GRCh37/hg19, February 2009) was performed using 'BWA' (Li and Durbin, 2009) or 'Bowtie2' (Langmead and Salzberg, 2012) under default settings. Duplicates were removed using Picardtools (version 1.60) and peak calling was carried out using MACS2 (version 2.1.1.20160309) setting a q value (FDR) to 0.05 or 0.01 (SMC1, SA1, SA2 in HMEC) and using the '--extsize' argument with the values obtained in the MACS2 'predictd' step (Zhang et al., 2008). All comparisons used the input tracks as "control", and each one of the datasets as "treatment".

Common, SA1-only and SA2-only positions were defined using BEDtools v2.26 (Quinlan and Hall, 2010)with a minimum of 1 nt overlap. Common positions were defined in two steps: 1) overlap between SMC1 and SA1 bed files was performed using '-wa -wb' argument and the positions obtained were concatenated and sorted using 'cat' and 'sort -k1,1 -k2,2n' commands. The output was merged using 'bedtools merge' function and considered as one dataset; 2) this was overlapped with the SA2 dataset as above. SA1-only and SA2-only positions are those where SA1 or SA2 do not overlap among each other.

Mean read density profiles and read density heatmaps for different chromatin binding proteins were generated with deepTools 2.0 (Ramírez et al., 2016) usam BAM files of processed reads and plotting them around peak summits of SA1 or SA2 only or common positions.

For Motif discovery analysis, whole sequences of cohesin positions were extracted and used for motif enrichment analysis using MEME-ChIP from MEME Suite (Bailey et al., 2009). Default parameters were used except for the following ones: -ccut 0, -meme-mod anr, -meme-minw: 6, -meme-maxw: 50, -nmeme: 600, -meme-nmotifs: 10, -meme-maxsize: 200,000.

Enrichment of cohesin positions (SA1 and SA2-only and common) at HMEC and HCAEC chromatin states (Ernst et al., 2011) was defined using 'intersect' function from BEDtools utilities (v2.26) with a minimum of 1nt overlap. The analysis was performed making sure that one position does not belong to two different chromatin states.

To analyse cohesin distribution along super-enhancers, ChIP-seq reads from SA1 and SA2 in HMECs and HCAECs were plotted along HMEC super-enhancers (Whyte et al., 2013) using the "scale-regions" parameter from deepTools to adjust all the super-enhancers to a predefined size and applying a local regression (LOESS) to smooth the read signals.

ChIP-qPCR and Re-ChIP. ChIP-qPCR on immunoprecipitated chromatin was performed using the SYBR Green PCR Master Mix and an ABI Prism® 7900HT instrument (Applied Biosystems®). Primers were designed using OligoPerfect Designer™ (Invitrogen) and reactions were performed in triplicate. Chromosome coordinates of the validated peaks and the corresponding primers are listed in Supplementary Table 7. The relative amount of each amplified fragment was normalized with respect to the amplification obtained from input DNA using the ΔΔCt method and represented as indicated in the corresponding figure legends.

ReChIP experiment was performed with the Re-ChIP-IT kit (#53016, Active Motif) according to the manufacturer's protocol. Briefly, MCF10A cells were fixed, lysed and sonicated as described in the ChIP protocol. Fifty ug of chromatin were incubated with 20 ug of the first

antibody (SA1, SA2 or IgG) in presence of magnetic beads, washed, eluted and further incubated with 5 ug of the second antibody (SA1, SA2, SMC1 or IgG). Eluted chromatin was analyzed by quantitative PCR. 1 ng of immunoprecipitated chromatin from two conditions, SA2 ChIP followed by IgG ReChIP and SA2 ChIP followed by SA1 ReChIP, was used to prepare libraries for Re-ChIP sequencing. Libraries were prepared with 18 PCR cycles. Peaks were called in SA2-SA1 ReChIP upon normalization with SA2-IgG ReChIP signals.

Quantitative RT-PCR (RT-qPCR) and RNA-sequencing (RNA-seq). cDNAs were prepared using the Superscript II reverse transcriptase (Invitrogen) from total RNA (RNeasy Mini Kit, Qiagen) and RT-qPCR analyses were performed using the SYBR Green PCR Master Mix and an ABI Prism® 7900HT instrument (Applied Biosystems®). Primers (Supplementary Table 7) were designed using OligoPerfect Designer™ (Invitrogen). Reactions were performed in triplicate. Quantifications were normalized to endogenous GAPDH, using the ΔΔCt method.

For RNA-seq libraries (three replicates for condition), polyA+RNA was purified with the Dynabeads mRNA purification kit (Invitrogen) from DNaseI-treated total RNA, randomly fragmented, converted to cDNA and processed through subsequent enzymatic treatments of end-repair, dA-tailing, and ligation to adapters as in Illumina's "TruSeq RNA Sample Preparation Guide" (Part # 15008136 Rev. A). Adapter-ligated library was completed by limited-cycle PCR with Illumina PE primers (8 cycles). The resulting purified cDNA library was applied to an Illumina flow cell for cluster generation (TruSeq cluster generation kit v5) and sequenced on the Genome Analyzer IIx with SBS TruSeq v5 reagents by following manufacturer's protocols. Fastq files with 50-nt single-end sequenced reads were quality-checked with FastQC (S. Andrews, http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and aligned to the human genome (GRCh37/hg19) with Nextpresso (http://bioinfo.cnio.es/nextpresso/) executing TopHat-2.0.0 using Bowtie 0.12.7 and Samtools 0.1.16 allowing two mismatches and five multi-hits. Transcript assembly, estimation of their abundances and differential expression were calculated with Cufflinks 1.3.0 using the human genome annotation data set from Ensembl. To account for multiple hypotheses testing, the estimated significance level (*p* value) was adjusted using Benjamini-Hochberg False Discovery Rate (FDR) correction. For differential expression, FDR<0.05, log2fold change<-0.5 or >0.5 and fpkm>3 in at least one of the two conditions compared was required.

GSEAPreranked was used to perform a gene set enrichment analysis (Subramanian et al., 2007). We used the RNA-seq gene list ranked by statistic, setting ‘gene set’ as the permutation method and we run it with 1000 permutations.

Hi-C. HiC was performed as described in (Rao et al., 2014) for *in-situ* HiC, with some modifications. Briefly: **Cell preparation and cross-linking:** 2-5 million of MCF10A cells were grown under above mentioned culture conditions and were arrested in G1 by means of high confluence (150,000 cells per cm^2). Cells were washed with PBS and a solution of 1% formaldehyde in serum-free medium was added directly to the plate. After 10 minutes of incubation at room temperature, glycine was added to a final concentration of 0.125 M to quench the formaldehyde and stop the cross-linking reaction. Cells were incubated for 5 minutes at room temperature and then put on ice for at least 15 minutes. Cells were washed with PBS and harvested in PBS using a scrapper. Restriction enzyme digestion step was performed using MboI enzyme. Two library replicates per condition were sequenced (>200 million read pairs each. Data were processed using TADbit (Serra et al., 2017) for read quality control, read mapping, interaction detection, interaction filtering, and matrix normalization. First, the reads were checked using an implemented FastQC protocol in TADbit. This allowed discarding problematic samples and detect systematic artifacts. Then, a fragment-based strategy in TADbit was used for mapping the remaining reads to the reference human genome (GRCh38). The mapping strategy resulted in about 80% of reads mapped uniquely to the genome. Next, non-informative contacts between two reads were filtered, including self-circles, dangling-ends, errors, random breaks or duplicates. The final interaction matrices resulted in 272 to 303 millions of valid interactions per experimental condition. These valid interactions were then used to generate genome-wide interaction maps at 100 Kb and 40 Kb to segment the genome into the so-called A/B compartments, Topologically Associating Domains (TADs), and produce differential interaction maps.

A/B compartments were calculated using vanilla normalized and decay corrected matrices as implemented in TADbit. Briefly, compartments are detected by calculating the first component of a PCA of chromosome-wide matrices and assigning A compartments to genomic bin with positive PCA1 values and high genes density (Fig. 6b). Conversely, B compartments are assigned to genomic bin with negative PCA1 values and low genes density. TADs were identified using 40 Kb resolution vanilla normalized and decay corrected matrices as input to the TAD detection algorithm implemented in TADbit. TAD border localization as well as strength was calculated and used to identify conserved borders and their strength (Fig. 6c-e). A border was considered conserved between siControl and siSA1 or siSA2 experiments if it was localized within +/- 2 two bins in both experiments. Raw matrices normalized by coverage (that is, all three experiments were scaled to have the same number of final valid interactions) at 100Kb resolution were also used for studying Hi-C interactions as function of genomic distance. This genomic decay was obtained per

chromosome to a maximum genomic distance of 50Mb and then averaged to obtain a genome-wide curve in siSA1 and siSA2 experiments (Fig. 6f). The same 100Kb matrices were used to determine differential Hi-C interactions between siControl and siSA1 or siSA2 experiments (Fig. 6g). These differential interactions maps were then used to assess the chromosome average differential interaction as a function of compartment localization and separating then in intra- and inter-TAD (Fig. 6h). Finally, the enrichment or depletion of genes (represented by their TSS), RNA (based on RNA-seq data), and CTCF and cohesin binding sites (SA1-only, SA2 and common) was analyzed by a log odds analysis of observing such features in genomic bins belonging to A, B compartments, A/B borders or TAD borders (Fig. 6i). The log odds distributions were assessed for their distribution being statistically different than zero as for a Fisher exact test (p-value <0.005).

Code availability. TADbit used for Hi-C analyses is freely available as a github repository at https://github.com/3DGenomes/tadbit

UCSC browser session with ChIP-seq data, chromatin states and HMEC super enhancers tracks is available using the following link: http://genome-euro.ucsc.edu/cgi-bin/hgTracks?hgS_doOtherUser=submit&hgS_otherUserName=Dinamica%20cromosomica&hgS_otherUserSessionName=Reviewers%20Session

Results

In the previous section, I have discussed that in addition to its canonical function in mediating sister chromatid cohesion, cohesin is essential for genome organization. Such important role is related with cohesin ring ability to entrap two DNA fragments “in cis”, resulting in the establishment of chromatin loops and TADs. Two versions of cohesin complex exist in somatic vertebrate cells, cohesin-SA1 and cohesin-SA2, and they are both present in all cell types. Even though no thorough dissection of the functional specificities of the two complexes has been conducted, work from our group and from other laboratories hinted to their non-redundant functions. Supporting the existence of functional specificities for both of cohesin variant complexes, in the last years it became clear the specific contribution of SA2 mutations in a number of human cancers.

Therefore, the main goal of this work is to explore the functional differences between cohesin-SA1 and cohesin-SA2 as well as to investigate the molecular basis underlying such specific functions. To this end, I have followed a functional genomics approach by combining the analysis of the genome-wide distribution of cohesin-SA1 and cohesin-SA2 across several non-tumorigenic human cell lines with the distribution of CTCF and number of histone modifications that define chromatin states in those same cell lines. The differential contribution of the two complexes to regulation and genomic organization was investigated by combining RNA-seq and 3C-derived techniques, in particular 4C and Hi-C in human mammary cells depleted from cohesin-SA1 or cohesin-SA2 by means of small interfering RNA (siRNA) technology.

1. Distinct genome-wide distribution cohesin-SA1 and cohesin-SA2

1.1. Relative abundance of cohesin variants in human cell lines from different origin

It has previously been reported that the relative abundance of cohesin-SA1 and cohesin-SA2 differs among cell types (Losada et al., 2000). In order to avoid biases due to unbalanced amounts of the two cohesin variants, we decided to select a cell line with comparable levels of both complexes to carry out our study. With this aim, we measured the levels of cohesin and the relative abundance of SA1 and SA2 in ten human cell lines representing different tissues and lineages by quantitative immunoblot (Figure R1).

We first compared the levels of total cohesin by measuring RAD21 signals and normalizing them to those of histone H3. Cohesin was most abundant in human mammary epithelial cells (HMEC), followed by cardiac endothelial cells (HCAEC) (Figure R1b). To determine the amounts of chromatin-bound SA1 and SA2, we performed a chromatin fractionation assay

for each cell line and assessed SA1 and SA2 levels by comparison with known amounts of purified C-terminal fragments of the two proteins (Figure R1c). Upon calculating and plotting the ratio between them, we observed that lung and mammary epithelial cells (NHBE and HMEC, respectively) were the cell lines that had the most comparable levels of SA1 and SA2 on chromatin (Figure R1d). Since HMEC also had more cohesin, we decided they were the most suitable cells for our study.

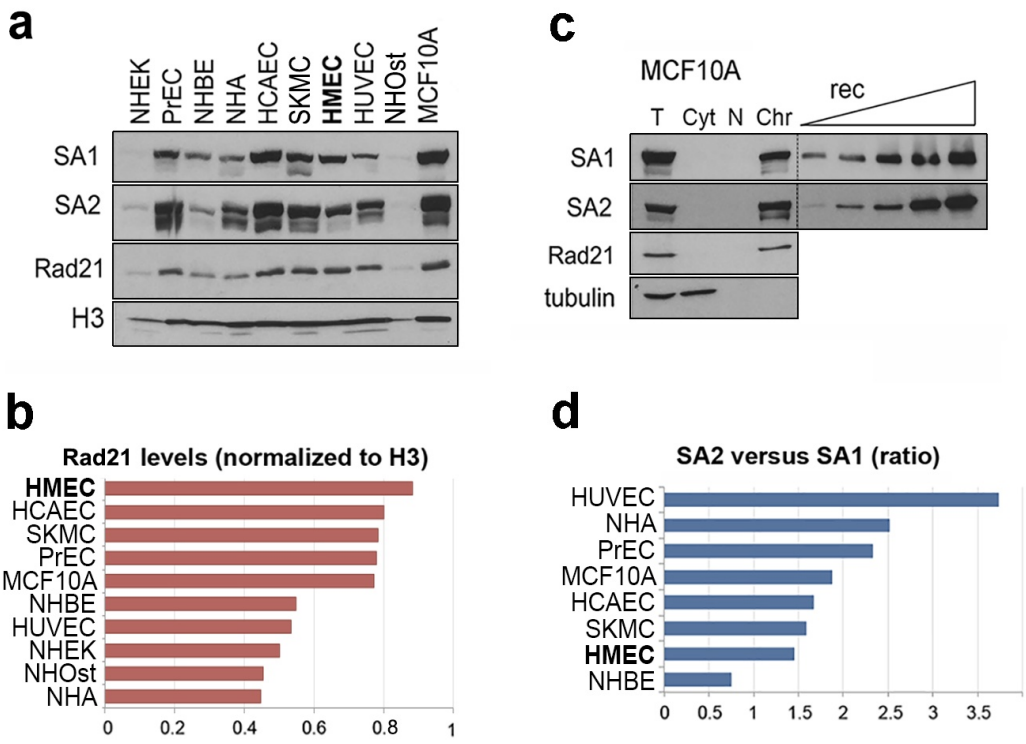


Figure R1. Cohesin variant abundance in different cell types

a. Immunoblot of total extracts from human primary cells (NHEK, Normal Human Epithelial Keratinocytes; PrEC, Prostate Epithelial Cells; NHBE, Normal Human Bronchial Epithelial Cells; NHA, Normal Human Astrocytes; HCAEC, Human Coronary Artery Endothelial Cells; SKMC, Skeletal Muscle Cells; HMEC, Human Mammary Epithelial Cells; HUVEC, Human Umbilical Vein Endothelial Cells; NHOst, Normal Human Osteoblasts) and MCF10A cells.

b. Quantification of the RAD21 signals (normalized to H3) for the blot shown in **a**.

c. Known amounts of C-terminal fragments of SA1 and SA2 (around 50 kDa each) were run alongside the indicated fractions from MCF10A cells to assess relative abundance of cohesin-SA1 and cohesin-SA2 on chromatin (T, total cell extract; Cyt, cytosolic fraction; N, Soluble Nuclear Fraction; Chr, Chromatin bound protein). The dotted line indicates that two parts of the film corresponding to the same exposure but different sizes have been pasted together. This was repeated with all cell lines used in A (not shown).

d. Histograms for SA2:SA1 ratio in the indicated cell lines.

1.2. Genomic distribution of cohesin variants in mammary epithelial cells identifies SA2-only positions lacking CTCF

Chromatin Immunoprecipitation coupled with deep sequencing (ChIP-seq) in HMEC was performed for the two SA isoforms as well as the common subunit SMC1 in order to warrant the specificity and reliability of the cohesin binding sites. We used custom made, validated antibodies and sequenced to high depth (around 100 million reads per antibody) to ensure whole genome coverage. Reads were aligned to hg19 version of human genome and peaks were called using MACS2 peak caller, setting FDR to <0.01. The number of peaks called for each antibody was: 45,985 for SA1, 82,077 for SA2 and 68,673 for SMC1. By overlapping called peaks using 'BEDtools intersect' utility we defined three different categories of cohesin positions: common (42,475), SA1-only (3,198) and SA2-only (39,061). Common cohesin positions had called peaks for both SA1 and SA2, whereas for SA1-only and SA2-only positions peaks were called for SA1 and not SA2 or SA2 and not SA1, respectively (Figure R2). These results showed that more than 90% of peaks called for SA1 are shared with SA2, while in contrast, in almost 50% of SA2 positions there is very little or no SA1.

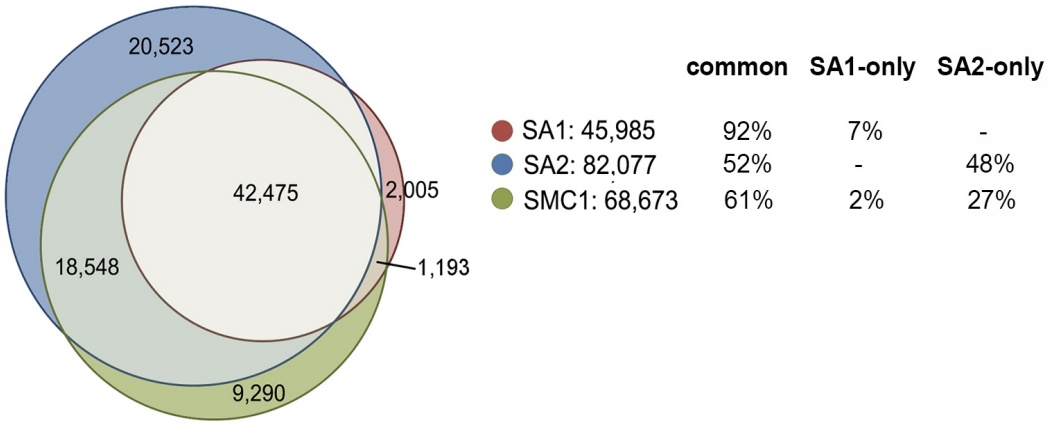


Figure R2. Three different categories of cohesin binding sites in HMEC

Venn diagram showing the overlap among the positions called for each cohesin subunit. On the right, percentage of positions in each of the three categories defined in the main text.

In order to better characterize these categories of cohesin positions and to check that SMC1 was present -even if the peak was not called- in all the SA1 and SA2 positions, we next generated heatmaps in which the reads from each experiment were plotted around the summits of the called peaks in the three categories as well as the corresponding read

density plots (Figure R3). Common cohesin positions were featured by high occupancy of cohesin and similar read density for SA1 and SA2. For both SA1-only and SA2-only positions, read density of SA1 and SA2 was lower than in the case of common positions. Importantly, while there was little SA1 in SA2-only positions, SA2 was still significantly enriched in SA1-only positions.

Next we asked about the presence CTCF in these positions for which we used data publicly available. Read density heatmaps and plots showed that CTCF was present in common cohesin positions and to a lesser extent, also in SA1-only positions. In contrast, SA2-only positions had undetectable levels of CTCF (Figure R3). Taken together, our data suggest that cohesin-SA1 function is largely related to that of CTCF, while cohesin-SA2 could have a unique, CTCF-independent function in approximately half of its positions.

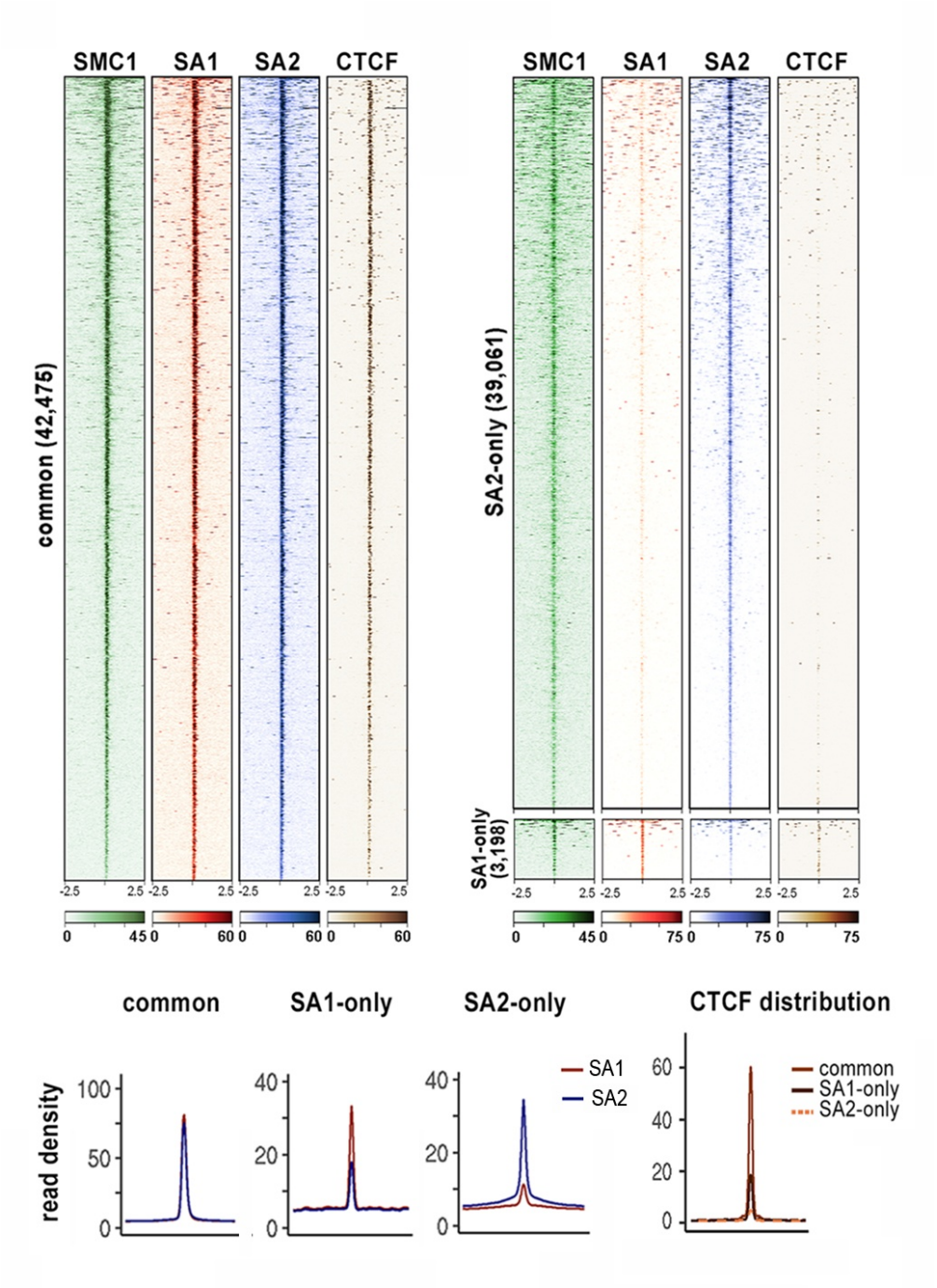


Figure R3. Three different categories of cohesin binding sites in HMECs

Analysis of ChIP-seq read distribution for SA1, SA2, SMC1 and CTCF around common, cohesin SA1-only and cohesin SA2-only positions within a 5-kb window in HMECs (top) and read density plots for SA1 (red) and SA2 (blue) distribution in common, SA1-only and SA2-only positions as well as for CTCF (separate plot on the right).

An additional advantage of the HMEC cell line is the availability of a broad number of genomic data, including histone marks such as H3K27Ac, H3K4Me3, H3K4Me1,

H3K27Me3, which permit the proper definition of chromatin states (Ernst et al., 2011). Overlap of cohesin positions with the chromatin states defined in HMEC revealed differential distribution of the two cohesin variants in the different genomic features. Around 40% of common positions were present at insulators, a state defined by the sole presence of CTCF, while 35% occurred within strong or weak enhancers (Figure R4). Some SA1-only positions were also found in insulators (23%) and enhancers (10%) but most were present in a chromatin state designated as “heterochromatin/low signal”, a poorly defined state likely related with silent or inaccessible genomic regions. Importantly, most (80%) of SA2-only positions were found in enhancers, particularly strong ones, and there was also a fraction in active promoters that was larger than the fraction of common sites in that state. This specific enrichment of SA2-only positions in active regions suggests a prominent role of SA2 in the regulation of gene expression.

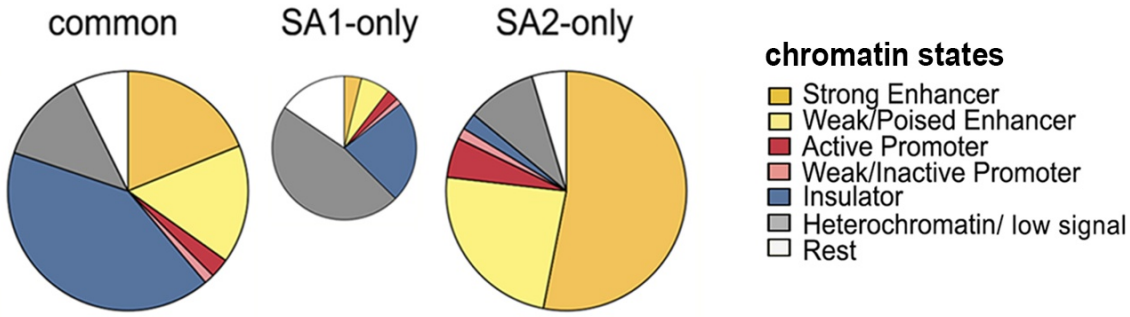


Figure R4. Most SA2-only cohesin positions are located in enhancers

Pie charts showing the distribution of cohesin positions in chromatin states defined in HMECs

Current models hold that cohesin binds DNA topologically without recognizing particular DNA motifs, but the role of CTCF as a boundary element stopping loop extruding cohesin leads to their colocalization at CTCF binding sites. The absence of CTCF at SA2-only positions prompted us to investigate the possible molecular determinants of CTCF-independent cohesin activities at these positions. We performed motif discovery analysis in all three categories of cohesin sites using “MEME-ChIP” utility within MEME suit (Bailey et al., 2009; http://meme-suite.org/). MEME is used for discovery of novel, ungapped DNA binding motifs within sequences of interest. MEME-ChIP (Machanick and Bailey, 2011) is a powerful tool for discovery and comprehensive analysis of DNA binding motifs performed on large sets of sequences coming from ChIP-seq experiments. In order to investigate all the transcription factors that could be located around cohesin positions we decided to extract the complete sequences recovered by ChIP-seq peaks (average size 400-500bp) and used

them as input for MEME. One thousand randomly chosen sequences were selected from each category of positions. Both common and SA1-only positions were significantly enriched in CTCF binding motif, an expected result given the presence of CTCF in most of these positions (Figure R3). In contrast, SA2-only positions were enriched in binding motifs for a number of different transcription factors including several members of the AP-1 family (Figure R5). In view of these data, we can conclude that while the function of cohesin-SA1 is exclusively and tightly related to CTCF, cohesin-SA2 could have specific and CTCF-independent functions, likely related to the regulation of transcription.

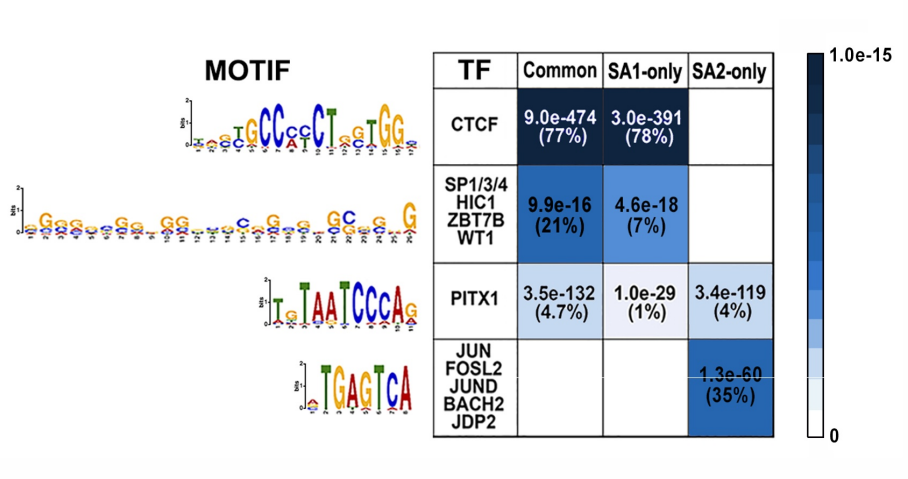


Figure R5. Transcription factor binding motives other than CTCF at SA2-only cohesin positions

Motif-based analysis of the indicated positions in HMEC cells. E-values of the significantly enriched motifs and percentage of regions containing each motif for a given condition are indicated. “TF” column contains the transcription factors with statistically significant binding (blue scale, on the right) to the identified motifs.

1.3. Identification of SA2-only positions in cardiac endothelial cells

To assess the general validity of our previous conclusions, we carried out analyses similar to those in HMECs in a cell line of different embryonic origin. We chose human cardiac endothelial cells (HCAEC) since, as HMECs, they displayed high levels of cohesin and similar SA1:SA2 ratio (Figure R1). With a sequencing depth of about 70 million reads per antibody (SMC1, SA1, SA2), the MACS2 algorithm and FDR<0.05, we obtained 40,796 and 39,072 called peaks for SA1 and SA2, respectively. This time we identified 30,034 common, 9,642 SA1-only and 7,942 SA2-only positions (Figure R6a). The read density profile plots demonstrated higher read density for SA1 than SA2 in common positions suggesting that the SA2 ChIP had been less efficient in these cells (Figure R6). We suspect that this has two consequences: first, that many SA2-only positions with low cohesin occupancy go

undetected and second, that a fraction of the positions defined as SA1-only are in fact common positions.

As in the case of HMEC, the levels of SA1 in SA1-only and of SA2 in SA2-only positions were lower than in common positions. Importantly, when we plotted the reads from CTCF distribution for a related cell line of endothelial origin (HUVEC, human umbilical vein endothelial cell), we confirmed that SA2-only positions lacked CTCF. Similar also to our findings in HMECs, SA2 could still be detected in SA1-only called positions, and these positions also showed high overlap with CTCF. Moreover, using the chromatin states defined for HUVEC (Ernst et al., 2011), we found enrichment of common positions in insulators (48%) and enhancer elements (36%) while SA2-only sites were mostly present at strong enhancers (65%) (Figure R6b). Thus, our analysis demonstrated that the presence of CTCF-independent, SA2-only cohesin positions located at enhancers is a common feature shared by different cell lines.

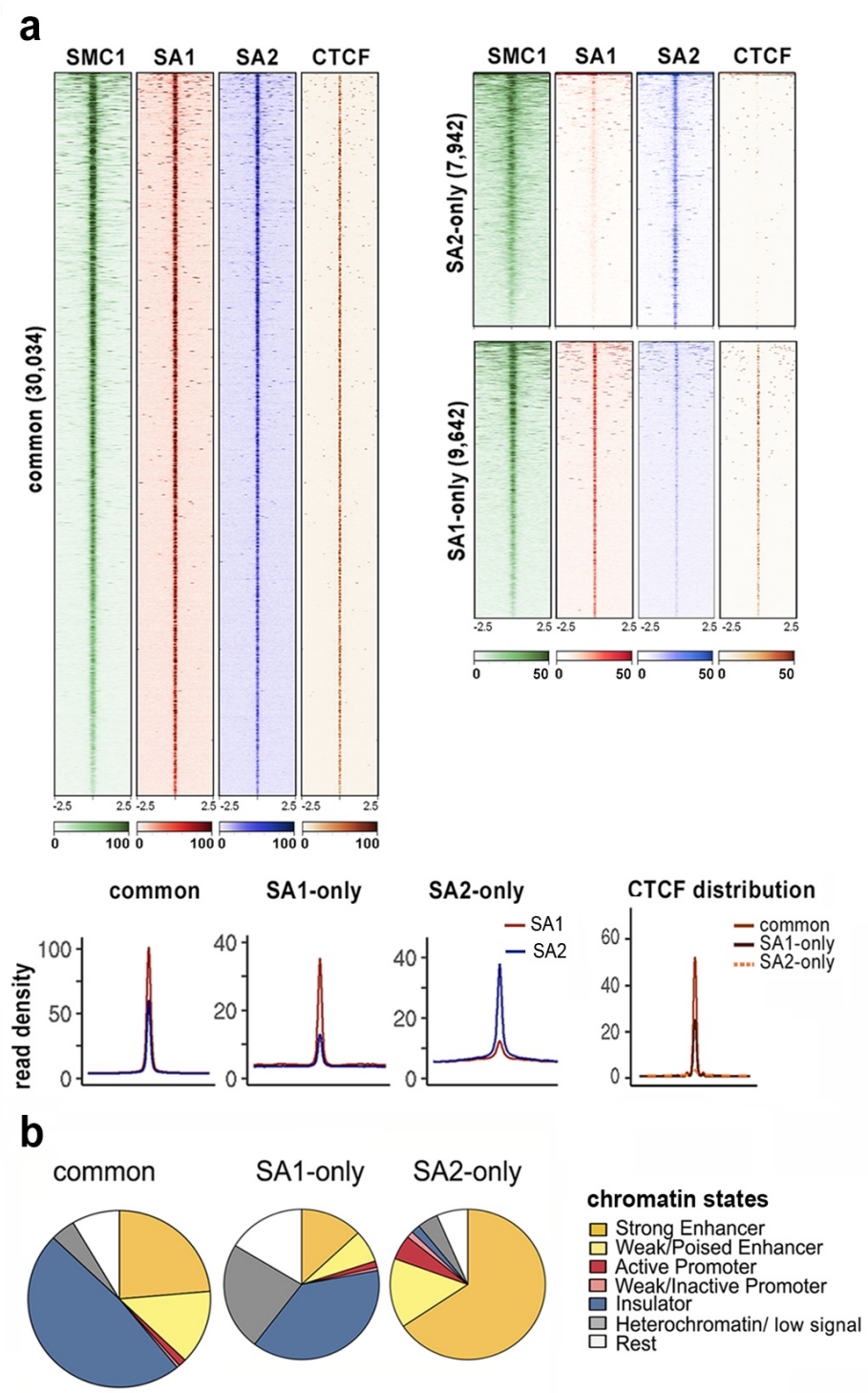


Figure R6. Identification of SA2-only positions at enhancers in cardiac endothelial cells

a. Analysis of ChIP-seq read distribution for SA1, SA2, SMC1 and CTCF around common, cohesin SA1-only and cohesin SA2-only positions within a 5-kb window in HCAECs (top) and read density plots for SA1 (red) and SA2 (blue) distribution in common, SA1-only and SA2-only positions as well as for CTCF (separate plot on the right).

b. Pie charts showing the distribution of cohesin positions in chromatin states defined in HUVEC.

A role for cohesin non-CTCF positions in the control of cell-type specific gene expression was suggested several years ago (Schmidt et al., 2010). In line with this, our analysis of cohesin distribution in two different cell lines suggests that SA2-only sites could have CTCF-independent cohesin roles. In order to test whether SA2-only positions are indeed cell-type specific, we overlapped cohesin positions defined in HMEC and HCAEC (Figure R7). While common positions were largely conserved between the two cell lines, this was not the case for SA2-only sites, confirming the cell-type specificity of this category of cohesin positions.

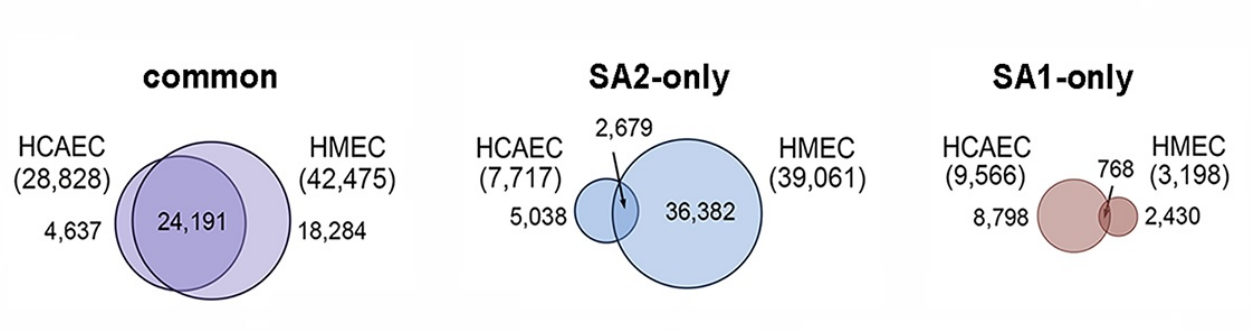


Figure R7. Transcription factor binding motives other than CTCF at SA2-only cohesin positions

Venn diagrams showing overlap of cohesin binding sites in the different categories between HMECs and HCAECs

In summary, our data demonstrate that cohesin-SA2 is the main variant at cohesin positions devoid of CTCF. These SA2-only positions are enriched at genomic regulatory elements, specially enhancers, containing DNA binding motifs for different transcription factors. We postulate that cohesin-SA2 is responsible for CTCF-independent roles of cohesin exerted from positions co-occupied by transcription factors other than CTCF. Moreover, we observed that SA2-only positions are variable between different cell lines, suggesting that these positions might be responsible for tissue-specific gene regulation. On the other hand, majority of cohesin-SA1 binding events do contain some SA2 and CTCF, and their distribution in chromatin states resembles that of common positions. Intriguingly, SA1-only positions are also quite different between HMEC and HCAEC, opening the possibility that such positions might play a functional role different to common ones.

1.4. Genomic distribution of cohesin variants in MCF10A cells

In order to further understand the functional specificities of the two variant complexes, we turned to a cell line that would be easier to grow and to transfect than the primary cell lines. We chose MCF10A, a transformed, non-tumorigenic cell line of human mammary epithelia

and first examined the genome-wide distribution of cohesin variants. ChIP-seq assays analysed as above yielded 50,299, 22,510 and 37,308 positions for SMC1, SA1 and SA2, respectively. By overlapping the called peaks we defined 20,154 common, 15,224 SA2-only and a very small number (1,489) of SA1-only positions. Read density of SA1 and SA2 in common positions was similar and the enrichment of CTCF around those sites was high (Figure R8). Most SA1-only positions also contained some CTCF and SA2, and its distribution in chromatin states was similar to that of common positions with a prevalence in insulators (Figure R9).

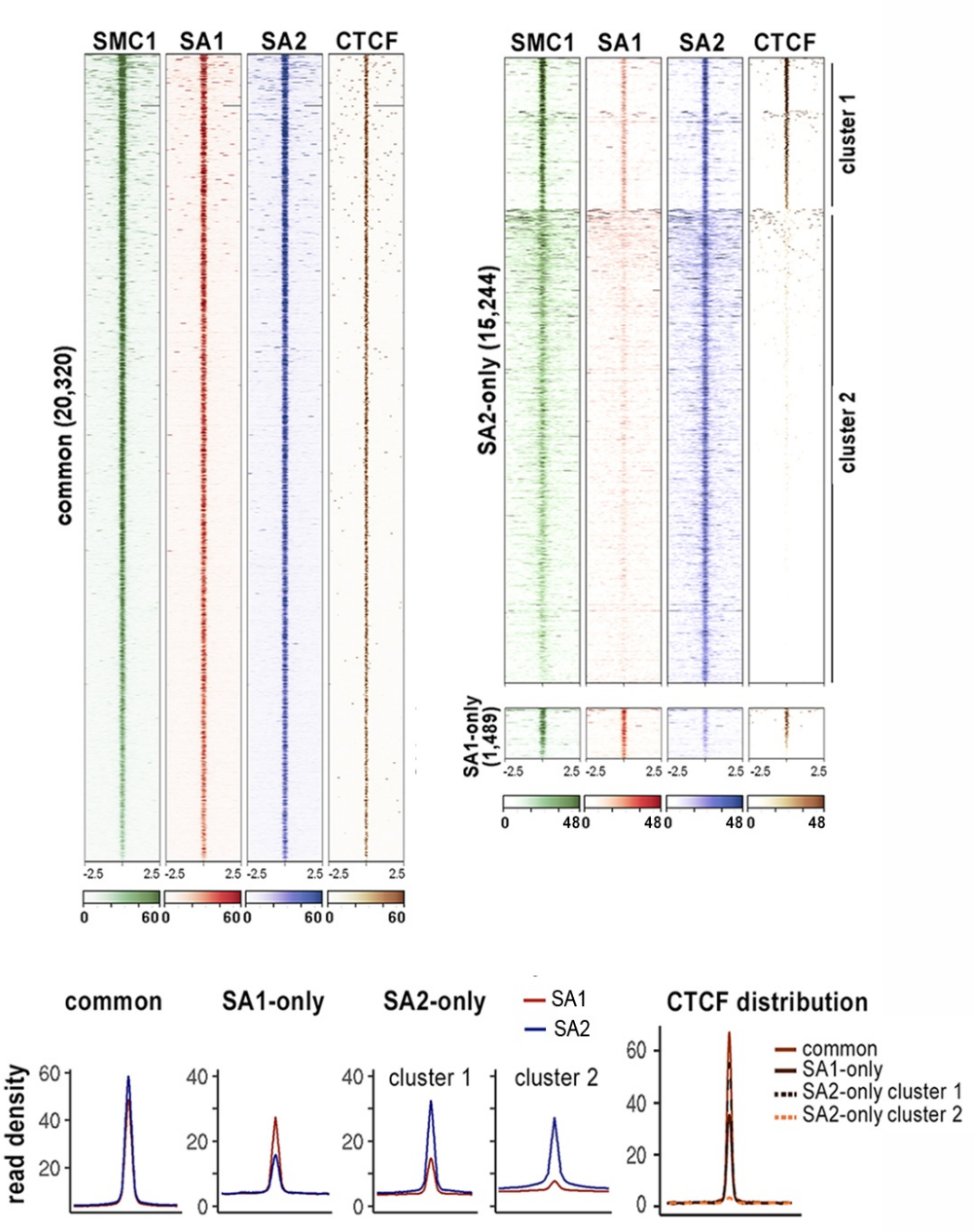


Figure R8. Three different categories of cohesin binding sites in HMECs

Analysis of ChIP-seq read distribution for SA1, SA2, SMC1 and CTCF around common, cohesin SA1-only and cohesin SA2-only positions within a 5-kb window in MCF10A cells (top) and read density plots for SA1 (red) and SA2 (blue) distribution in common, SA1-only and SA2-only positions as well as for CTCF (separate plot on the right). SA2-only position have been subdivided in two clusters. Cluster 1 groups positions that are most likely common positions in which SA1 ChIP efficiency was lower.

Among SA2-only positions, read distribution heatmaps distinguished two clusters. Cohesin positions in both clusters were enriched in SA2, but those in cluster 1 contained also some SA1 and CTCF and its distribution among chromatin states was not very different to that of common positions (Figure R9). The larger cluster 2, in contrast, comprised SA2-only positions without CTCF and were enriched in enhancers and depleted in insulators compared with common and SA1-only positions. Taken together these analysis support that MCF10A cells constitute an appropriate cell model for our next experiments.

Figure R9. Most SA2-only cohesin positions are located in enhancers

Distribution of cohesin positions identified in MCF10A in chromatin states defined in HMECs.

1.5. Cohesin-SA2 interacts with transcriptional regulators

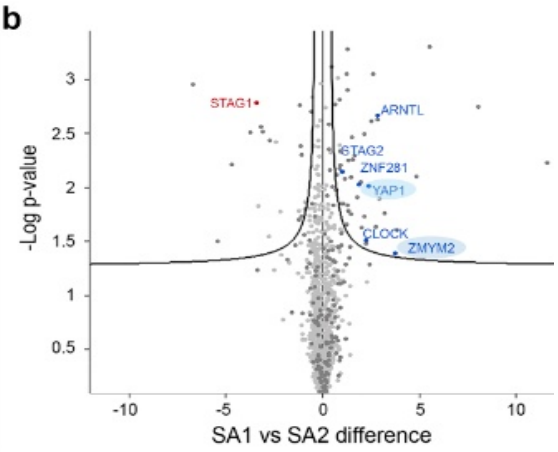
Aimed to identify specific partners of cohesin-SA2 responsible for CTCF-independent cohesin functions, we performed mass-spectrometry analysis of immunoprecipitates obtained from MCF10A cell extracts with either SA1 or SA2 antibodies, and IgG as control. Cohesin subunits and its regulators could be identified in both cases (Figure R10a), but a number of transcriptional regulators were detected only in the SA2 pull down, including ZMYM2 and YAP1 (Figure R10b). ZMYM2 function is related with transcriptional repression through its association with the CoREST complex while YAP1 is a transcriptional co-activator. To corroborate these results, we examined the genomic distribution of ZMYM2 in MCF10A by ChIP-seq and obtained 1,243 peaks. The heatmap and read density profiles of SA1, SA2 and CTCF centred at ZMYM2 peak summits demonstrated that there is a clear preference for SA2 at those positions, while CTCF and SA1 are absent. Likewise, re-analysis of a ChIP-seq dataset for activated YAP1 from another study (Jang et al., 2017) according to our parameters yielded 9,393 peaks that, as in the case of ZMYM2, showed specific enrichment of SA2-only positions (Figure R10c).

The data from the proteomic analysis together with the specific enrichment of SA2 positions around the transcriptional regulators ZMYM2 and YAP1 strongly suggest that cohesin-SA2 is responsible for CTCF-independent cohesin functions through its association with transcription factors.

a

FUNCTION	Protein	MS counts IgG	SA1	SA2
COHESIN subunits	SMC1A	0	14	82
	SMC3	2	18	41
	RAD21	4	52	136
	SA2	0	11	43
	SA1	0	48	0
COHESIN Associated Factors	PDS5B	4	11	48
	PDS5A	0	1	6
	WAPL	0	4	43

b



c

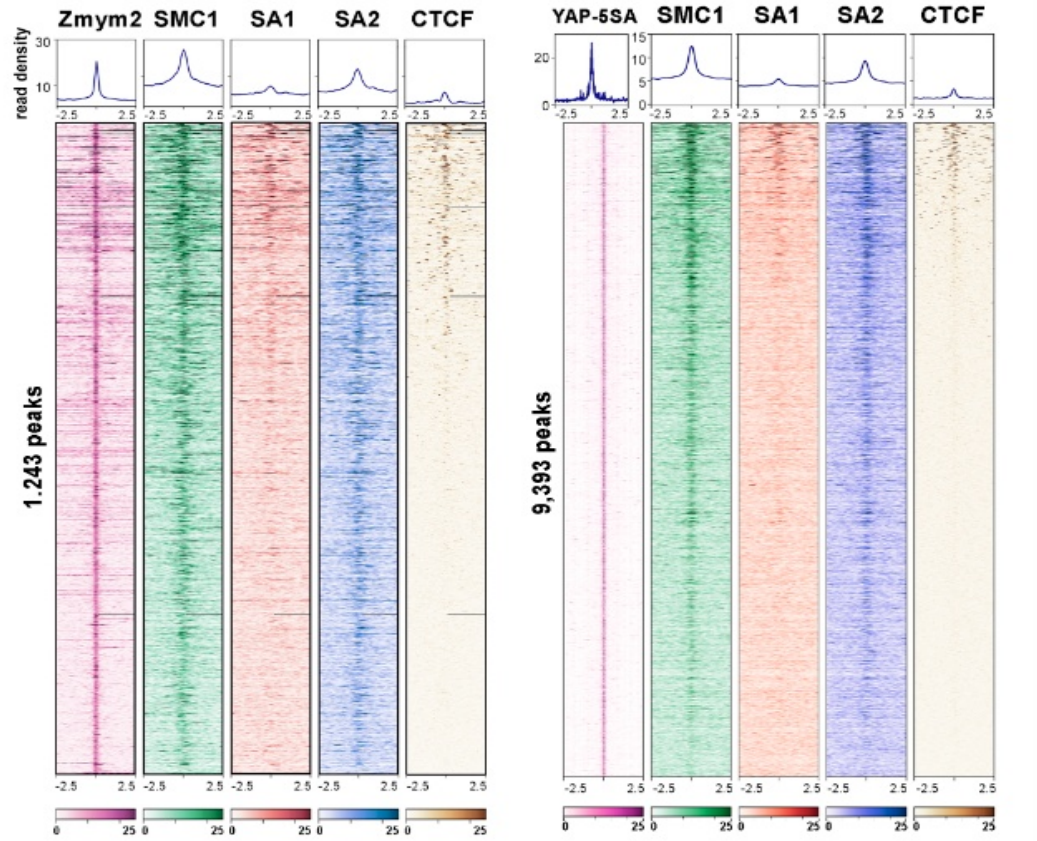


Figure R10. Cohesin-SA2 interacts with transcriptional regulators

a. Table showing th mass spectrometry (MS) counts obtained for cohesin and its regulators in the immunoprecipitates obtained from MCF10A cell extracts with antibodies against SA1 and SA2 but not IgG (control).

b. Identification of significant differential interactions for SA1 and SA2. Dark grey dots show proteins significantly enriched compared with the mock immunoprecipitation using IgG. Proteins whose function is related with transcriptional regulation are indicated in blue.

c. Distribution of SMC1, SA1, SA2 and CTCF around ZMYM2 and activated YAP (YAP-5SA) positions (± 2.5 kb) defined by ChIP-seq in MCF10A cells.

2. Different dynamics of cohesin-SA1 and cohesin-SA2 association with chromatin

We wished to understand the molecular determinants responsible for the different genomic distribution and relationship with CTCF and other chromatin binding proteins exhibited by both cohesin variants. The experiments reported in this section aimed to explore differences in their dynamic association with chromatin that could account for their functional specificities.

2.1. Cohesin complex is more dynamic in SA2-only than in common positions

Close visual inspection of SMC1 read density plots, revealed that while the distribution of SMC1 reads around common and SA1-only positions produces sharp and narrow profiles in the three cell lines under study, such distribution around SA2-only positions gives much broader profiles (Figure R11). Consistently, SA1 and SA2 profiles around common positions are also sharp and narrow. This observation suggests that cohesin bound at SA2-only positions is more dynamic than cohesin bound at common or SA1-only sites.

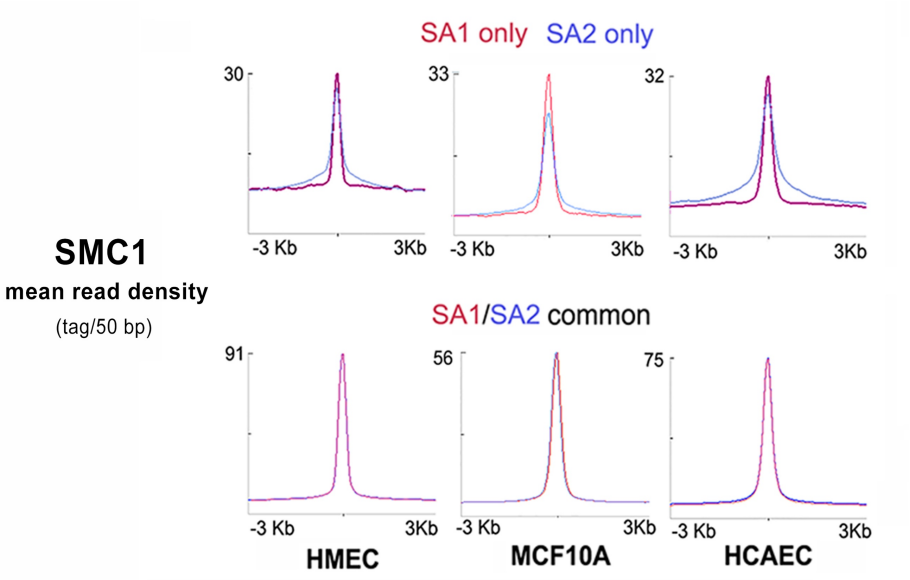


Figure R11. Different behavior of cohesin at common and SA2-only positions

Average read density plots for SMC1 ChIP-seq reads in the SA1-only, SA2-only (top) and common (bottom) cohesin positions identified in the three human cell lines indicated.

A plausible explanation to this behaviour is that cohesin at common positions is less prone to be removed from chromatin by the activity of WAPL when compared to cohesin at SA2-only positions. To test this hypothesis, we measured the levels of SA2 and WAPL at several common and SA2-only positions from by ChIP-qPCR (Figure R12a). SA2/WAPL ratio was much higher at common positions than SA2-only positions. This suggests that cohesin at common positions is less likely to associate with WAPL and be removed by its activity compared to SA2-only positions.

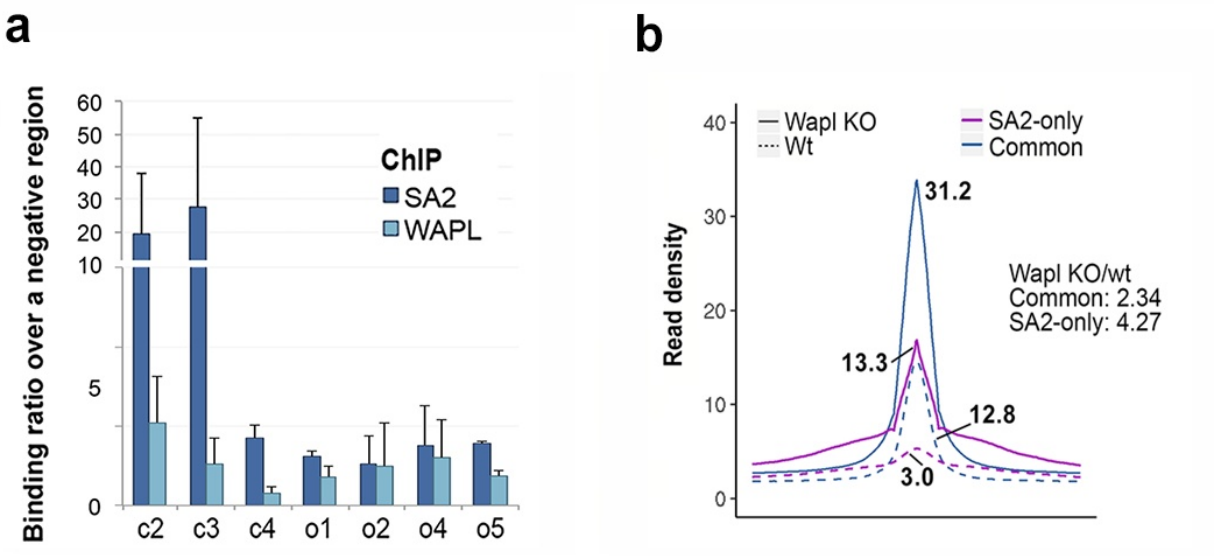


Figure R12. Cohesin is less likely to associate with WAPL at common positions

a. SA2 and WAPL binding to three common (c2,c3,c4) and four SA2-only positions (o1, o2, o4, o5) from the S100 gene cluster described in next section were assayed by ChIP-qPCR. Bars represent the mean of at least three independent experiments performed in triplicates; error bar=SD

b. Read density plots showing cohesin SMC1 distribution in wild type (wt) and WAPL knockout (KO) human HAP1 cells. Cohesin positions were classified into common and SA2-only based on their co-occurrence with CTCF in WAPL KO cells.

If this idea were correct, one would expect that elimination of WAPL would increase the amount of chromatin bound cohesin at SA2-only positions more than at common positions. A recent study in human leukemic HAP1 cells aimed to analyse the consequences of WAPL removal on chromatin loop dynamics provided data on SMC1 and CTCF distribution in wild type and WAPL KO cells (Haarhuis et al., 2017). We decided to test our prediction with these data. First, to identify common and SA2-only positions, we re-analysed ChIP-seq data and assumed that those cohesin positions that lack CTCF were actually SA2-only positions, while those positions in which cohesin colocalized with CTCF were common positions. When we checked the read density of SMC1 in both conditions we found that the increase of SMC1 at SA2-only positions was almost double the increase at common positions (Figure R12b). These findings support the idea that WAPL removes more efficiently cohesin bound at SA2-only positions. Interestingly, a mass-spectrometry analysis performed to identify SA1 and SA2 specific interactors revealed stronger preference of WAPL for cohesin complexes containing SA2 (Figure R10a).

2.2. Cohesin-SA2 is less stably bound to chromatin than cohesin-SA1

After observing that binding of cohesin is more dynamic in SA2-only positions than in common positions we decided to further explore the stability of chromatin association of the two SA variants using a salt extraction experiment. To this end, the chromatin fraction of MCF10A cells was treated with two different concentrations of salt (0.25M or 0.5M NaCl) for either 10 or 20 min and the amount of each variant remaining on chromatin was assessed by immunoblotting. We found that SA2 was more sensitive to the salt than SA1, as seen at all time points in the lower salt treatment. At higher salt the enhanced sensitivity of SA2 could be seen at the shorter time point (Figure R13). We conclude that the association of cohesin-SA2 with chromatin is less tight or, in other words, more dynamic, than the association of cohesin-SA1. It is possible that the higher resistance of cohesin-SA1 to salt stems from its preferential association with CTCF.

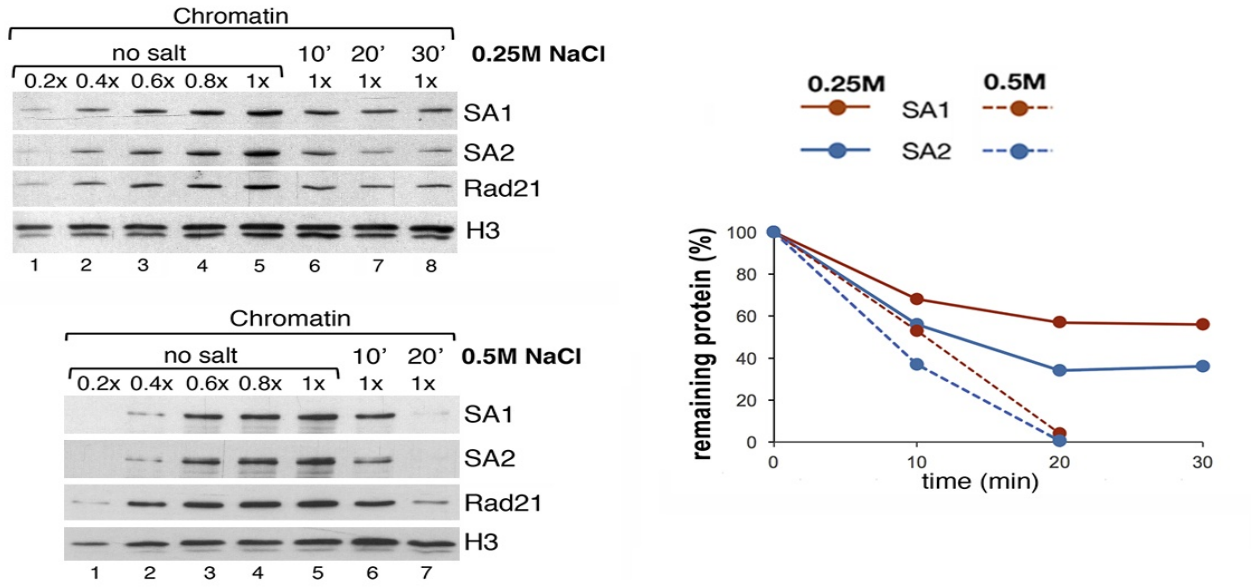


Figure R13. Cohesin-SA2 binds less tightly to chromatin than cohesin-SA1

Remaining cohesin on chromatin after treatment of chromatin fractions with 0.25M and 0.5M NaCl for the indicated times was assayed by quantitative immunoblotting. Quantification is shown on the right.

2.3. More than one cohesin complex can coexist in the same position in a single cell

One important question in the field is whether more than one cohesin complex can occupy the same position in a given cell. The fact that cohesin complexes carrying one of two distinct subunits, SA1 or SA2, can be found by ChIP at the common positions, gave us the opportunity to address this question using Re-ChIP. This assay involves two sequential rounds of immunoprecipitation, using antibodies specific for the two proteins whose colocalization is to be tested. We evaluated cohesin enrichment at five common positions and two "negative" regions selected based on our ChIP-seq data in MCF10A cells. After the first immunoprecipitation, we observed specific enrichment of both cohesin variants at common positions, as expected, but not at negative regions, while no binding was detected with the IgG (Figure R14, upper panel). After elution of chromatin from antibodies and thorough washes, the second round of immunoprecipitations was performed with SA2 in SA1 immunoprecipitated chromatin (second panel from top) and with SA1 in SA2-immunoprecipitated chromatin (third panel from the top). As control, both were also immunprecipitated with SMC1 and IgG as positive and negative controls, respectively. The latter Re-ChIP sets the background level, since it reflects the antibody carry over from the first ChIP. One further control was to immunoprecipitate the chromatin eluted from IgG beads

in the first immunoprecipitation with SA1 or SA2 antibodies, which gave no significant enrichment in any of the studied regions (Figure R14, lower panel).

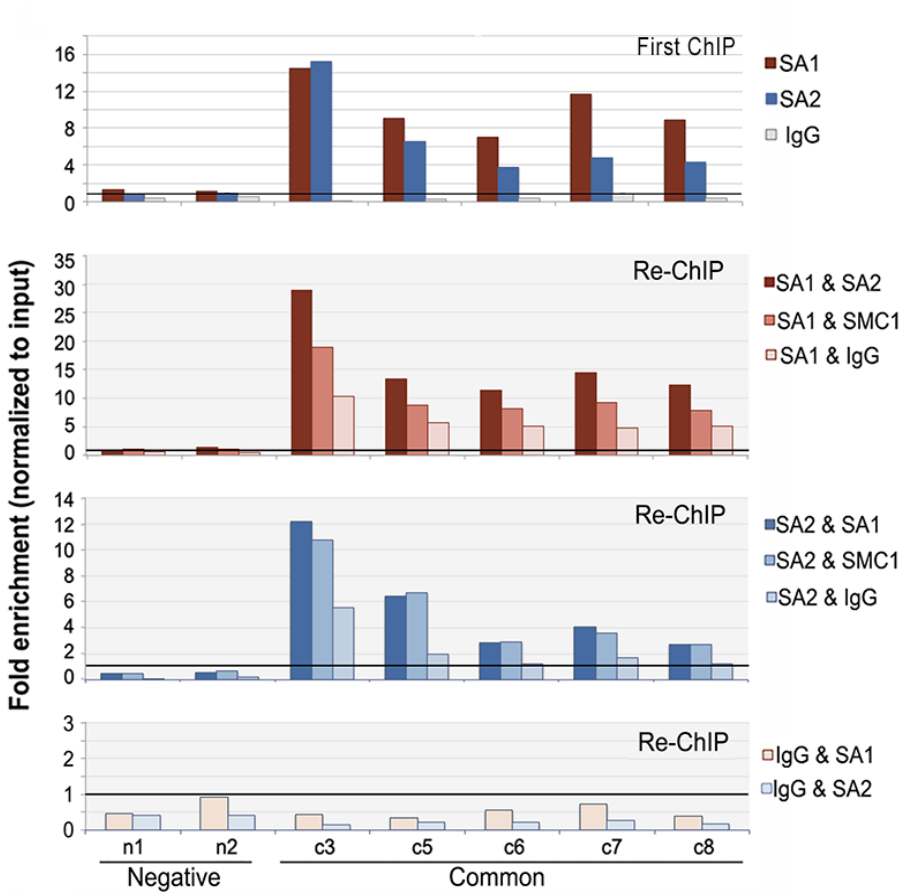


Figure R14. Two cohesin complex can coexist at the same position

Re-ChIP assay. Chromatin eluted from the first ChIP with SA1 or SA2 was incubated with SA2 and SA1 antibodies, respectively, as well as SMC1 and IgG as positive and negative controls. Lower panel, Re-ChIP of chromatin eluted from IgG beads with SA1 and SA2. Positions c3-c8 are "common" cohesin binding sites; n1 and n2 are negative regions.

Upon removal of the first antibody (SA1 or SA2) and incubation with the second antibody (SA2 or SA1 respectively), we could detect cohesin enrichment over the IgG background in all the cohesin binding sites tested (second and third panels from top in Figure R14). The incubation of SA1 or SA2 immunoprecipitates with anti SMC1 as the second antibody confirmed cohesin enrichment.

In an attempt to further explore the genomic regions where both cohesin isoforms coexist simultaneously, we generated ChIP-seq libraries from some of the samples of the Re-ChIP experiment, SA2&SA1 and SA2&IgG. Preparing these libraries was specially challenging,

due to the limited amount of available material. In spite of the technical limitations, upon using SA2&IgG signal as the "input" in the peak-calling step, we obtained 201 called positions in the SA2&SA1 Re-ChIP condition. Heat map density plots of SA1, SA2 and CTCF centered on those Re-ChIP positions and its distribution in chromatin states showed that all of them correspond with common sites (Figure R15).

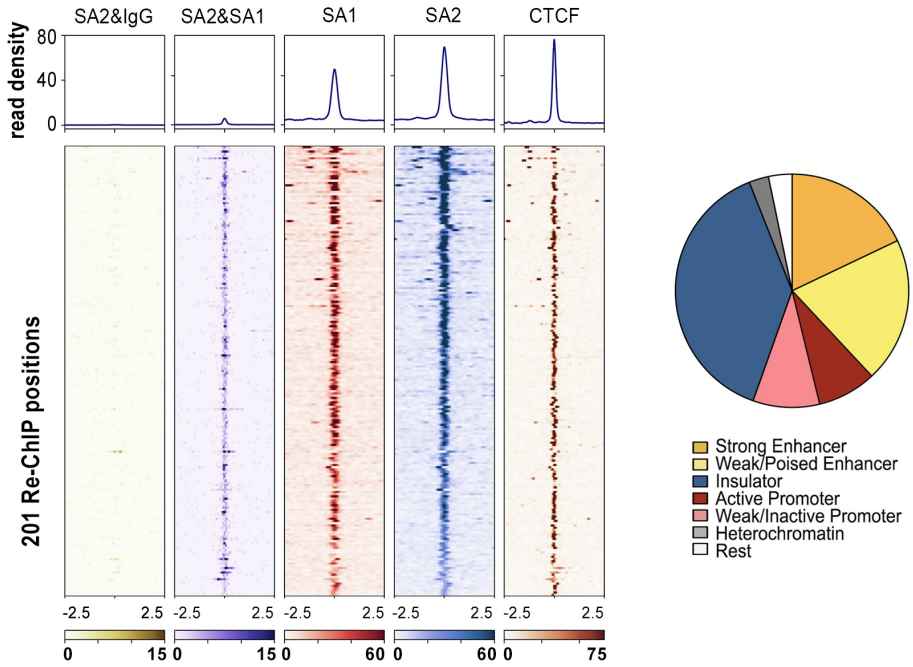


Figure R15. Genome-wide assessment of sites occupied by two cohesin complexes

Heatmaps showing the distribution of the indicated proteins around the 201 positions obtained after ChIP-seq analyses of the chromatin obtained by Re-ChIP of SA2 and SA1 (top) and distribution of these positions in chromatin states.

Altogether, these data suggest that at least two different cohesin complexes can simultaneously be present in the same genomic location within the same cell.

2.4. Cohesin-SA1 cannot occupy SA2-only sites in SA2-depleted cells

To understand if cohesin-SA1 and cohesin-SA2 can replace each other and therefore assume their respective specific functions in situations where their relative levels change, we studied their genomic distribution upon SA1 or SA2 depletion using calibrated ChIP-seq. Calibrated ChIP-seq involves mixing the sample under study with a little amount (20% in our experiments) of chromatin from a different species, the "calibration" sample, whose sequences can be easily distinguished during the aligning process, and in which

experimental steps of fixation, sonication and Immunoprecipitation work in a similar fashion. An important requirement is that the protein expressed by both the calibration and the experimental species is recognized by the antibody used for immunoprecipitation. The “calibration” genome is used as an internal control for the experiment, which helps to more accurately compare samples immunoprecipitated with the same antibody under different conditions (Hu et al., 2015). MCF10A cells were transfected with small interfering RNAs (siRNAs) specific for SA1 or SA2 or mock transfected as control and downregulation of each protein was assessed by immunoblotting (Figure R16).

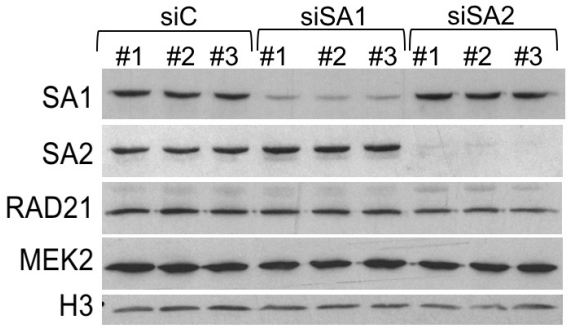


Figure R16. Downregulation of SA1 and SA2 by in MCF10A cells by siRNA

Cell were mock transfected (siC) or transfected with siRNA against SA1 (siSA1) and SA2 (siSA2). After 72h they were collected and the depletion of SA1 and SA2 was checked by immunoblotting. Transfections were carried out in triplicates with similar efficiency.

ChIP was performed in duplicate with SA1 and SA2 antibodies in mock, SA1 and SA2 depleted cells. Around 20% and 80% of the total number of reads obtained in each condition aligned to the mouse and human genomes, consistent with the ratio in which mouse and human chromatin were mixed (see Tablexx in Appendix). We first generated heatmaps with the reads that aligned to the mouse genome to show the degree of technical variation among replicates and conditions (Figure MM1).

Next we generated heatmaps and read density profiles for SA1 and SA2 in the three conditions and centered them at summits of previously defined common, SA1-only and SA2-only positions in MCF10A cells (Figure R18). We observed that, in SA1 depleted cells, there was little cohesin-SA1 left at any position while the presence of cohesin-SA2 increased both at common and SA2-only sites compared to mock transfected cells. Importantly, in SA2 depleted cells, cohesin-SA1 could not occupy cohesin-SA2 only sites and instead accumulated further at common positions. We conclude that both cohesin-SA1 and cohesin-SA2 can reach common binding sites independently of each other. Thus, in the absence of

one variant, the other could at least partially compensate for its loss at these CTCF bound sites. In contrast, cohesin-SA1 cannot occupy cohesin-SA2 only positions when SA2 is missing. This suggests that cohesin functions dependent on those positions could be severely impaired in SA2-depleted cells.

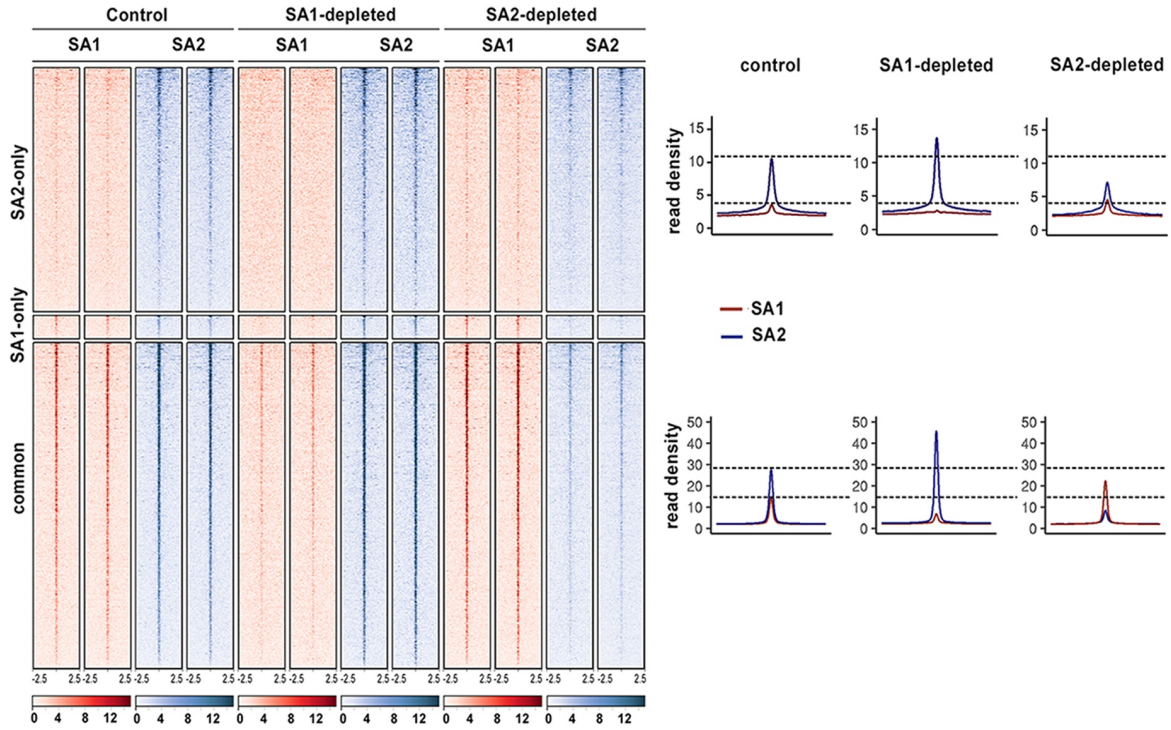


Figure R18. Cohesin-SA1 cannot replace cohesin-SA2 at SA2-only positions

Heatmaps (left) and read density plots (right) showing SA1 and SA2 distribution around cohesin positions defined in Figure R8 in control cells, SA1 depleted cells and SA2 depleted cells. Two independent replicates were performed per condition. Read density plots were built merging the reads from the two replicates.

3. Differential contribution of cohesin variants to the control of gene expression

3.1. Depletion of SA1 or SA2 has different impact on the transcriptome of MCF10A cells

We have demonstrated that the existence of different categories of cohesin positions is a general feature of all cell lines. Unpublished results from our group confirm their presence in mouse embryonic stem cells and fibroblasts. While common cohesin-SA1 and cohesin-SA2 positions are conserved in different cell types and largely overlap with CTCF, SA2-only positions are cell-type specific, CTCF-free, and are located in enhancers and active

promoters together with transcription factors. Based on these results, it is tempting to propose that cohesin isoforms might have different roles in gene expression, being SA2 the most relevant cohesin in the control of cell-type specific transcription. To test this hypothesis, we decided to analyse the changes in the transcriptome of MCF10A cells upon depletion of SA1, SA2 or CTCF using RNA-seq.

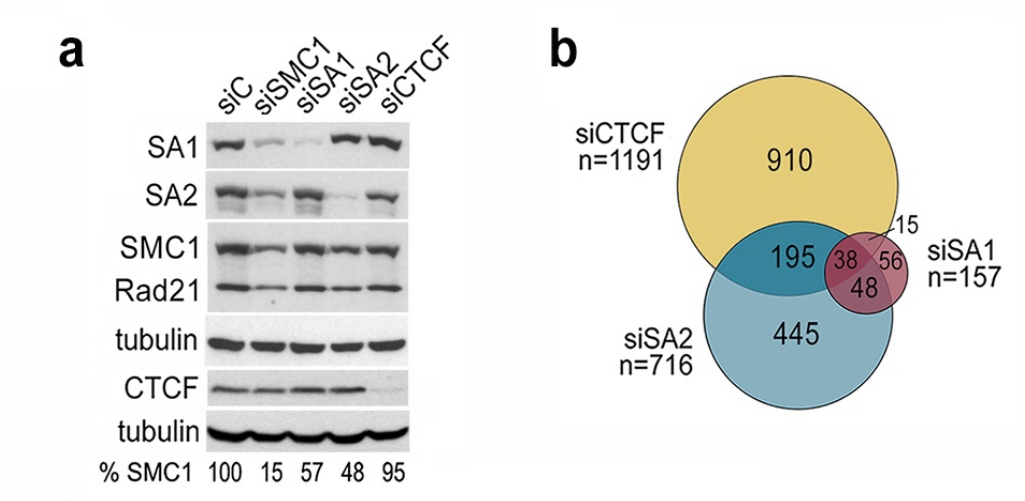


Figure R19. Most SA2-only cohesin positions are located in enhancers

a. Levels of cohesin and CTCF after siRNA transfection in MCF10A cells. Cells were also mock transfected (siC) and transfected with siRNA against SMC1 (siSMC1) for comparison. The amount of cohesin (SMC1) remaining in each condition is shown below. Tubulin is shown as loading control.

b. Venn diagram showing the overlap between genes deregulated in the indicated condition compared to mock transfected cells (FDR<0.05, log_2fold change <-0.5 or >0.5 and FPKM >3 in at least one condition). Experiments for siSA1 and siSA2 were performed in triplicate where for CTCF to replicates were analysed.

We transfected MCF10A cells with small interfering RNAs (siRNAs) specific for SA1, SA2 and CTCF. Cells were collected 72 hours after transfection and the remaining protein levels were assessed by immunoblot (Figure R19a). The efficiency of SA1 and SA2 depletion was comparable and, importantly, decreased expression of each of the SA variants led to a similar reduction of total cohesin amount, as indicated by SMC1 levels. As previously described, depletion of CTCF, had no impact on cohesin. RNA was extracted from the cells and analysed and libraries were prepared for sequencing.

In order to ensure significant and meaningful transcriptional changes, we applied the following criteria to our RNA-seq analysis: FPKM>3 in at least one of the two compared conditions and FDR<0.05 and log_2 fold change <-0.5 or >0.5 for differential expression. As a

result, we obtained 157 and 716 differentially expressed genes (DEGs) upon depletion of SA1 and SA2 respectively. This number raised to 1,191 in the case of CTCF depleted cells (Figure R19b). Interestingly, as many as 493 genes SA2-dependent genes were not affected by CTCF depletion, confirming a prominent CTCF-independent role for SA2 in the regulation of transcription. On the other hand, most DEGs found in SA1 depleted cells (101 out of 157) were also found to be deregulated upon depletion of SA2 or CTCF suggesting that, indeed, the transcriptional role of cohesin-SA1 is tightly linked to the activity of these two proteins.

3.2. Cohesin-SA2 contributes to maintain cell-type identity

We have just shown that depletion of SA2 has stronger impact on transcription than depletion of SA1 and that this effect is largely CTCF independent. Aimed to identify the consequences of these transcriptional changes and the biological processes affected, we performed Gene Set Enrichment Analysis (GSEA) with the transcriptomic data from SA1, SA2 and CTCF depleted cells. Significant upregulation of pathways related to the nervous system and the hematopoietic cell lineage was observed specifically upon SA2 depletion (Figure R20a). This result could reflect a certain degree of loss of cell identity under this condition

Individual inspection of DEGs in SA2 depleted cells further revealed abnormal expression of genes known to be important tissue-specific transcription factors. In particular, one of the genes that appeared on top of the list of genes upregulated in siSA2 cells was BDNF (Brain Derived Neurotrophic Factor), an important neuronal transcription factor. Of note, BDNF expression is known to be repressed in non-neuronal tissues by the activity of CoREST complex (Ballas et al., 2005). In addition, Irx3 and Tfap2c, two core transcription factors in mammary gland cells (D'Alessio et al., 2015) were found down-regulated upon SA2 depletion. These findings suggest that cohesin-SA2 is contributing to maintain cell-type identity by regulating the expression of tissue-specific core regulators.

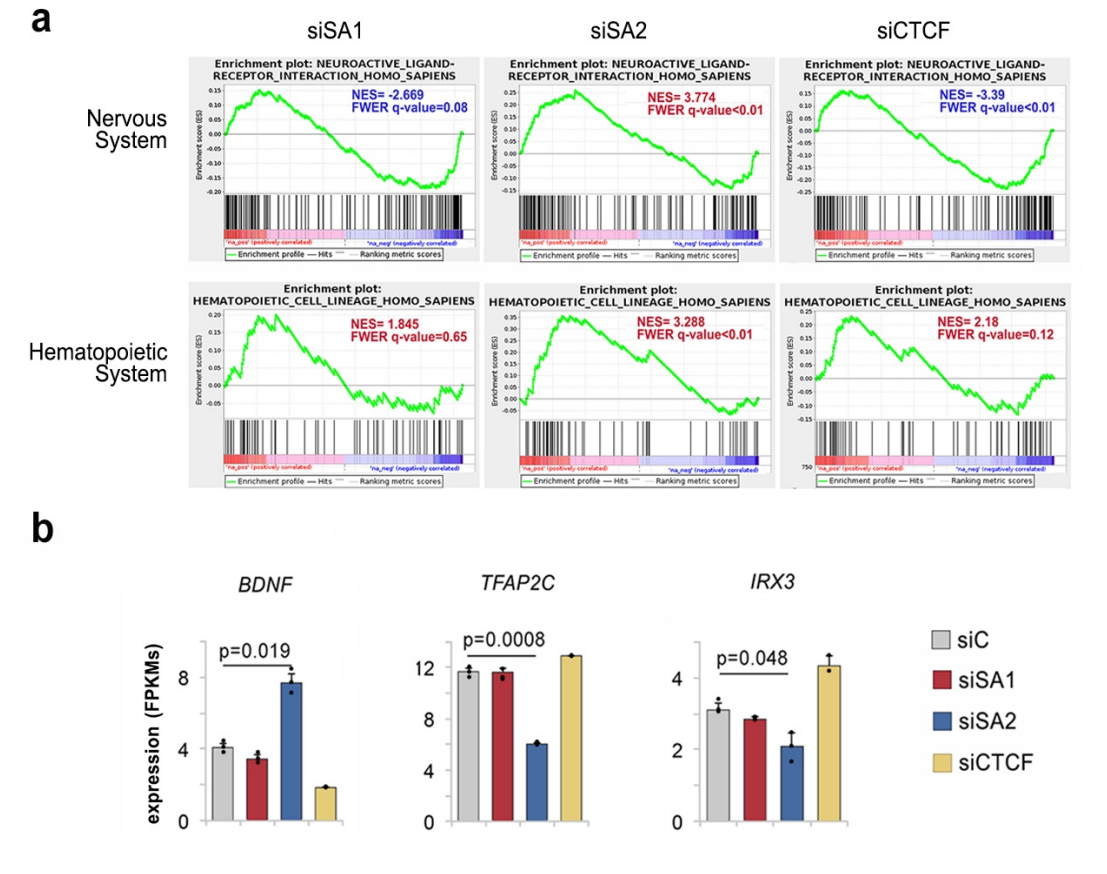


Figure R20. Cohesin-SA2 contributes to maintain cell identity

a. Enrichment plots for genes within KEGG pathways specific for the hematopoietic system ("Hematopoietic cell lineage") and the nervous system ("Neuroactive Ligand-Receptor interaction") after downregulation of SA1, SA2 or CTCF. Both these pathways appear significantly upregulated in the siSA2 condition (NES>0, in red; FWER q-value<0.01). In the other two conditions the gene expression changes are either not significant (FWER q-value in grey) or in the opposite direction (NES<0, in blue). Y axis, enrichment score (ES).

b. Gene expression levels (in FPKMs) of cell-type specific transcription factors in control (siC), siSA1, siSA2 and siCTCF conditions from the RNA-seq analysis. Student's T test was used to assess statistical significance.

To further confirm this hypothesis, we examined the relative enrichment of the two cohesin SA isoforms in the super-enhancers (SEs), which play an essential role in the control of mammalian cell identity (Hnisz et al., 2013). These regulatory elements consist of clusters of enhancers that are densely occupied by cell-type specific master regulators and the Mediator complex. SEs differ from typical enhancers in size, transcription factor binding density and content, ability to activate transcription and sensitivity to perturbation. They were initially discovered in ESCs as regulatory elements present at most genes that control the pluripotent state, densely bound by ESC master transcription factors, but their existence and activity has also been confirmed in differentiated cells (Whyte et al., 2013). Using the list of

SEs defined for HMECs, we first performed a simple overlap between the three categories of cohesin positions defined in this cell line and the genomic regions containing SEs. We observed that these regions are specially enriched in SA2-only positions (Figure R21a), and that this enrichment is cell type specific.

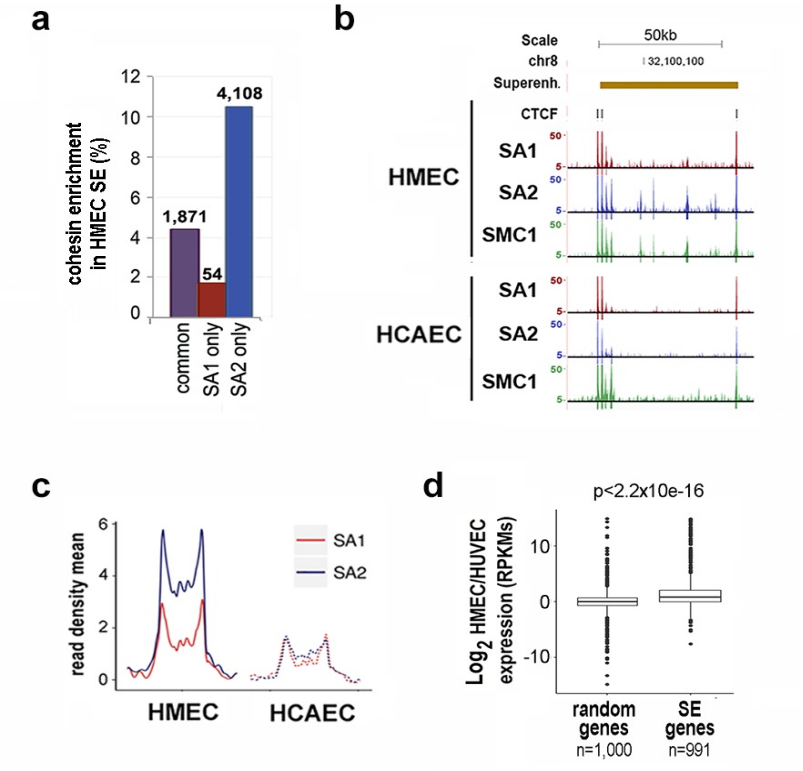


Figure R21. Cohesin-SA2 is enriched at super-enhancers

a. Cohesin enrichment in super-enhancers (SE) defined in HMECs

b. Snapshot of the browser showing distribution of cohesin around an HMEC SE in HEMCs and HCAECs.

c. Average read density plot showing SA1 and SA2 enrichment in HMEC and HCAECs along all HMEC SE.

d. Box-plot that compares changes in expression between random genes and genes associated with HMEC-specific SE. Statistical significance was calculated with a Wilcoxon signed-rank test. Data from Whyte et al., 2013.

In Figure R21b a HMEC SE is shown that is flanked by common cohesin positions with bound CTCF and contains additional cohesin-SA2 only peaks lacking CTCF along its more than 50kb long region. Importantly, the SA2-only positions are present in HMECs, when the SE is active, but not in HCAECs. Average read-density profiles for SA1 and SA2 ChIP-seq reads from the two cell lines along all HMEC SEs confirmed that cohesin enrichment, specially cohesin-SA2, was much lower in HCAEC than in HMEC (Figure R21c). Moreover,

increased gene expression levels of the genes potentially regulated by SEs in HMEC is found in HMECs compared to the endothelial HUVEC cells while this is not the case in a control group of randomly selected genes (Figure R21d). Taken together, these results support a role for cohesin-SA2 in cell-type specific transcription.

3.3. Further insights in the role of cohesin-SA2 in gene expression: The S100 cluster

By close inspection of our RNA-seq data we noticed that a number genes encoding members of the S100 protein family were significantly downregulated specifically in SA2-depleted MCF10A cells. S100 proteins regulate a plethora of intra- and extra-cellular processes and altered expression of many of them is associated to cancer progression and poor prognosis (Chen et al., 2014). Interestingly, 22 genes of this family are distributed within a 300-kb cluster located in chromosome 1q21. When we examined cohesin distribution in MCF10A cells in the region containing S100 cluster we observed a number of prominent common positions that co-localized with CTCF as well as several SA2-only positions featured by lower read density and lack of CTCF overlap. Interestingly, most of these SA2-only positions were located precisely at the promoters of deregulated S100 genes (orange dots in Figure R22a).

We selected four common (c1 to c4) and three SA2-only (o1 to o3) positions and validated our ChIP-seq results by quantitative ChIP-qPCR (Figure R22b). We also confirmed by quantitative RT-qPCR the specific downregulation of S100 genes after SA2 depletion detected by RNA-seq (Figure R22c). Importantly, depletion of SA1 did not have any impact on the expression of these genes.

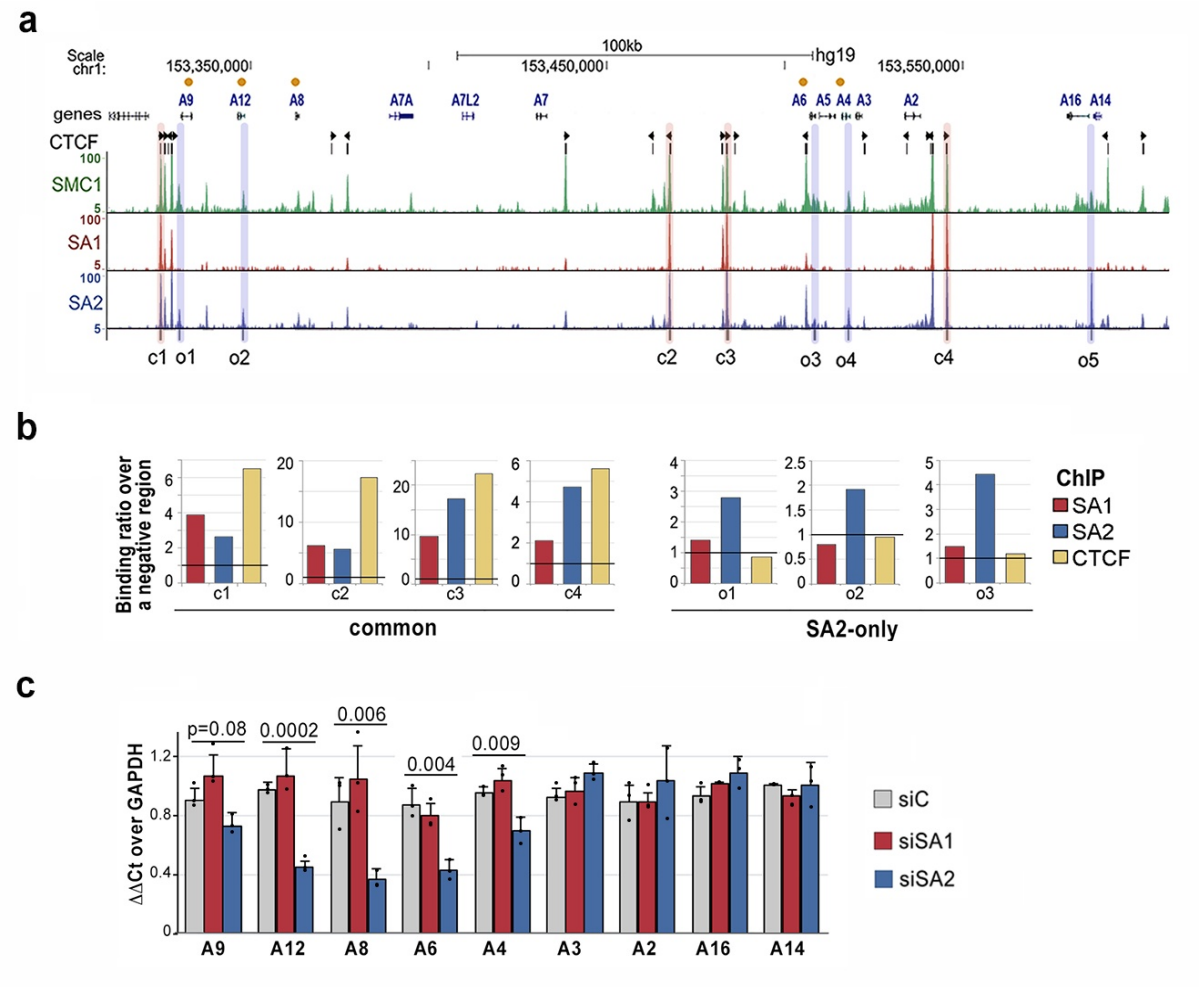


Figure R22. Cohesin-SA2 is enriched at super-enhancers

a. UCSC browser image of the S100A gene cluster showing genes (orange dots indicate those deregulated in siSA2 cells), CTCF peaks and distribution of SMC1, SA1 and SA2 in MCF10A cells. Some common (c) and SA2-only (o) positions are shadowed in red and blue, respectively.

b. ChIP-qPCR validation of SA1, SA2 and CTCF binding to the indicated (c) and (o) positions.

c. Gene expression levels of the indicated S100 genes assessed by RT-qPCR (n=3 experiments). Student's T test was used to assess statistical significance.

We next asked if the genomic distribution of cohesin-SA1 and cohesin-SA2 along the S100 cluster would contribute to the architecture of the locus. To answer this question, we first looked at the spatial organization of the S100A locus using publicly available Hi-C data from epithelial (HMEC) and endothelial (HUVEC) cells (Rao et al., 2014) and observed striking differences between both cell lines (Figure R23a).

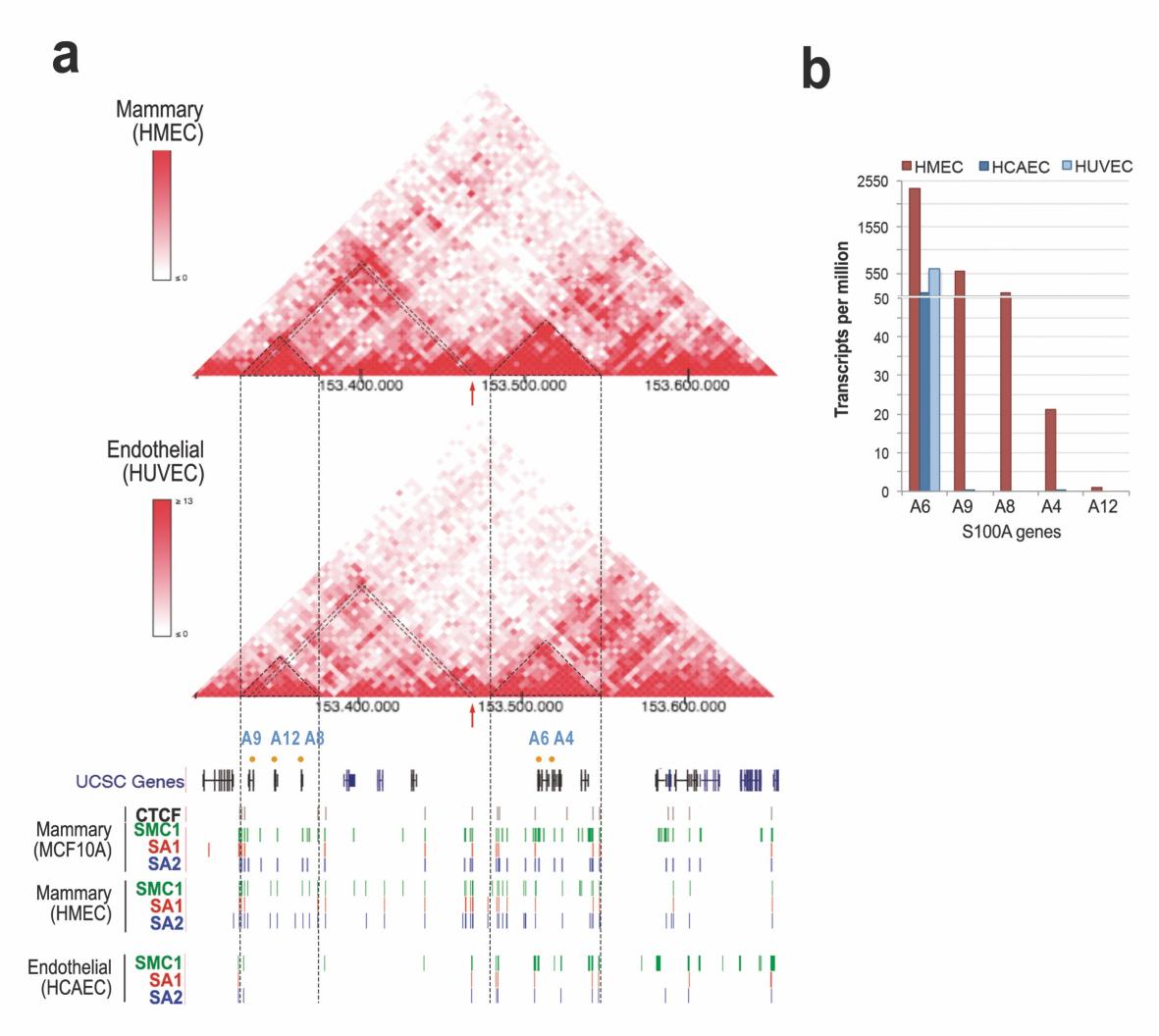


Figure R23. Cohesin distribution correlates with genome organization and gene expression

a. HiC contact matrices for S100A locus at 5kb resolution in mammary (HMEC) and endothelial (HUVEC) cells is shown. The intensity of each pixel represents the normalized number of contacts between a pair of loci. Positions for cohesin subunits and CTCF are indicated below.

b. Expression levels of the genes of the S100A locus deregulated in siSA2 MCF10A cells (orange dots in a) in HMEC, HUVEC and HCAEC.

In HMECs we observe frequent and strong contacts within each of the two isolated domains that encompass this locus (highlighted by dotted lines). In contrast, these interactions are much reduced in endothelial cells, coinciding with the absence of SA2-only positions specifically in this last cell type. Thus, there seems to be a correlation between the strength of local interactions within a contact domain and the presence of cohesin SA2-only positions. Importantly, the expression levels of the S100A genes located within the strongly interacting domains, are also much higher in mammary than in endothelial cells (Figure R23b).

Cohesin and CTCF have been shown to demarcate contact domains. Since cohesin-CTCF sites contain both cohesin-SA1 and cohesin-SA2, we hypothesized that elimination of SA2 might have no major impact on this demarcation function. To test this hypothesis, we devised a 4C experiment. As explained in the Introduction, 4C is a 3C-derived technique that allows the identification of all genomic regions that interact with a defined position or “viewpoint”. We chose as viewpoint one of two strong common positions the two subdomains of the locus (arrow in Figure R23a) and examined its interaction profiles in three conditions: mock depleted cells as control and cells depleted from SA2 or CTCF. In the control cells, all the interactions of the locus happen within the subdomain on the right (red curves in Figure R24). When CTCF is depleted, reduced insulation allows for new interactions with the regions to the right of the viewpoint to arise (blue curves in Figure R24). In contrast, and in line with our hypothesis, SA2 depletion does not have a visible impact in insulation.

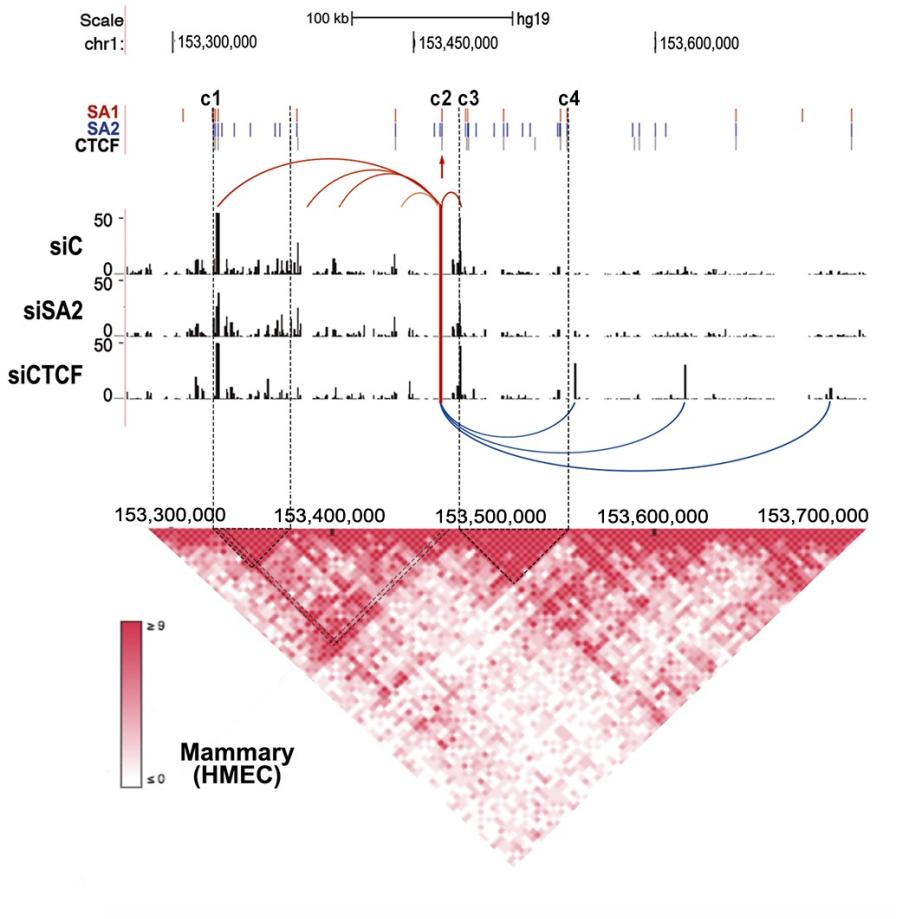


Figure R24. SA2 depletion does not reduce insulation

A 4C experiment in S100A locus was performed in MCF10A in mock depleted cells (siC) wand cells with reduced SA2 (sSA2) or CTCF (siCTCF) levels. A vertical red line and an arrowhead indicate the viewpoint position. Binding sites of cohesin-SA1, cohesin-SA2 and CTCF identified by ChIP-seq are also shown. Red curves highlight interactions present in the control cells. Blue curves indicate new interactions detected after CTCF depletion. The Hi-C contact matrix from S100A locus at 5kb resolution in HMEC cells from Figure R24 is shown for reference.

Altogether, our data suggest that at least in the context of S100A cluster, SA2 does not play a major role in maintaining the overall locus structure. However, there is a strong correlation between the presence of SA2-only positions, the existence of local interactions and the levels of expression of nearby genes.

4. Hi-C analyses show that cohesin-SA1 and cohesin-SA2 make different contributions to genome architecture

Our previous results demonstrate differences in SA1 and SA2 cohesin variants regarding their genome-wide genomic distribution, their relationship with CTCF and a number of

transcription factors, their functional enrichment in chromatin states and their dynamics on chromatin. These differences probably account for their distinct impact on gene expression, which is very likely related with the existence of specific roles of both cohesin variants in the 3D organization of the genome.

To address the specific contributions of cohesin isoforms to genome architecture, we carried out Hi-C experiments in MCF10A cells depleted from SA1 or SA2 by siRNA. In order to avoid the effect of cell cycle on possible differences in the interaction profiles, we enriched the cell populations in G1 by growing cells at high confluency to arrest cell cycle by contact inhibition. Depletion levels estimated by immunoblot were similarly high for SA1 and SA2 (Figure R25a) and between 70 and 80% of the cells were in G1 (Figure R25b).

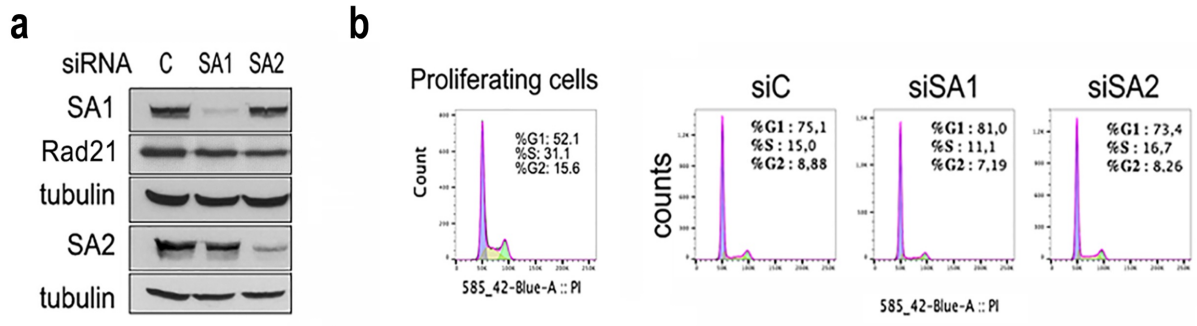


Figure R25. Cells depleted from SA1 or SA2 by RNAi and enriched in G1 for Hi-C analyses

a. Immunoblot analysis to assay the remaining levels of cohesin 72h after transfection of the indicated siRNAs. Tubulin is used as loading control.

b. Cell cycle profiles of MCF10A proliferating cells and of each of the samples used for Hi-C analysis. Percentages of cells in each phase are indicated.

We generated two libraries per condition from two replicates for each biotin and sequenced around 250 million read pairs per replicate. Data analyses were carried out by Marc Marti Renom at CNAG-CRG and details are given in the Methods section. Details on sequencing of the HiC experiment are summarized in Table 5 of the ANNEX.

Normalized Hi-C contact maps for all chromosomes were generated at 100Kb resolution for each replicate (Figure 26a). They showed high correlation between replicates and the same was true for Eigen values (Figure R26b and R26c). For analysis described below, only results obtained merging reads from both replicates of each condition are shown, but in all cases individual replicates were also analysed.

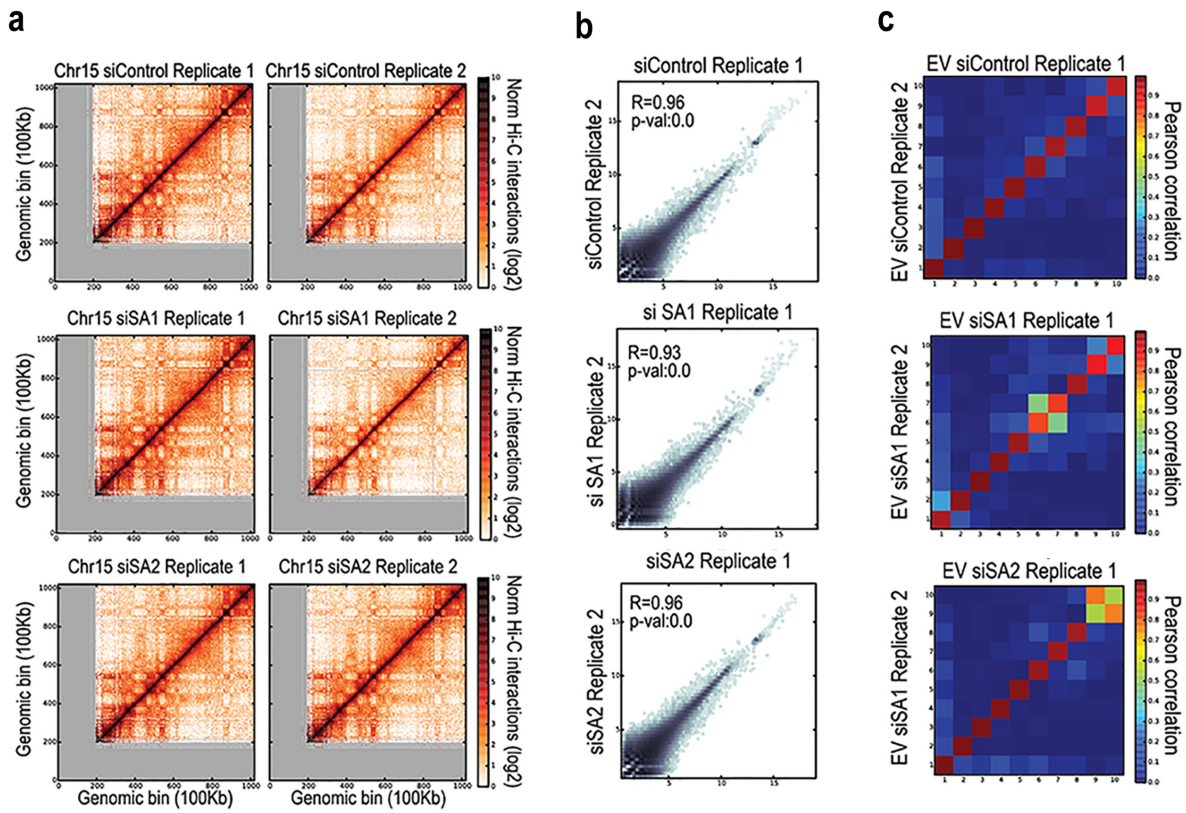


Figure R26. Hi-C analyses of MCF10A cells with reduced levels of SA1 or SA2

a. Hi-C normalized interaction maps for chromosome 15 at 100Kb resolution for the two replicates (r1 and r2) for each condition siC, siSA and siSA2.

b. Correlation between replicates of each condition of all data in the normalized interaction map for chromosome 15

c. Correlation between the Eigen values.

4.1. Role of cohesin variants on TAD and compartment conservation

We first examined whether depletion of SA1 or SA2 affected the identity of active euchromatic (A) and repressive heterochromatic (B) compartments. Visual inspection of the Eigen values from which the analysis of compartments is done for individual chromosomes shows no clear alterations in compartmentalization (Figure R27a). Genome wide, only a small percentage of bins changed from A to B, or vice versa (Figure R27b). In other words, the identity of compartments was largely unperturbed with depletion of SA1 or SA2.

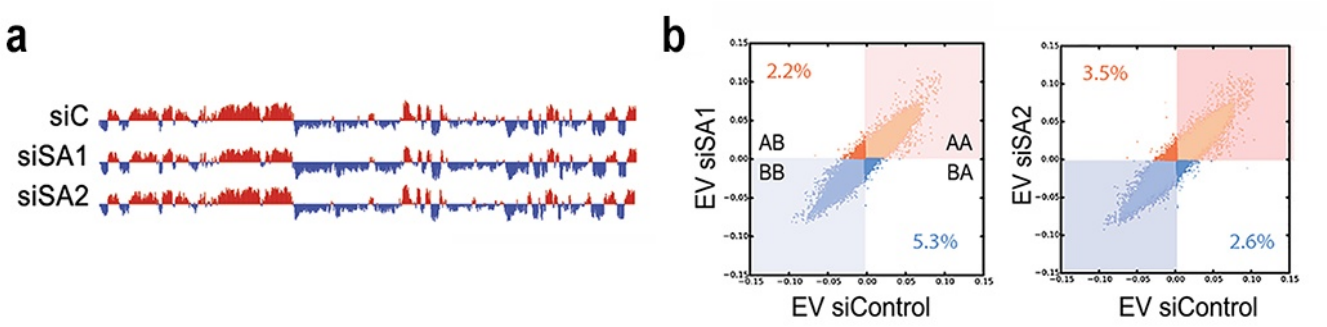


Figure R27. No change in compartment identity in the absence of SA1 or SA2

a. First eigenvector for chromosome 13 at 100Kb resolution. A compartment, red; B compartment, blue.

b. Scatter plot of eigenvectors (EV) of the intra-chromosomal interaction matrices indicated. Numbers within the plot show the % of bins that change compartment.

Then we asked to which extent depletion of one or the other cohesin variant affected overall segmentation of genome into topologically associated domains or TADs. At first glance, contact maps for the three conditions appear to be similar (Figure R28). However, a more detailed analyses revealed some differences.

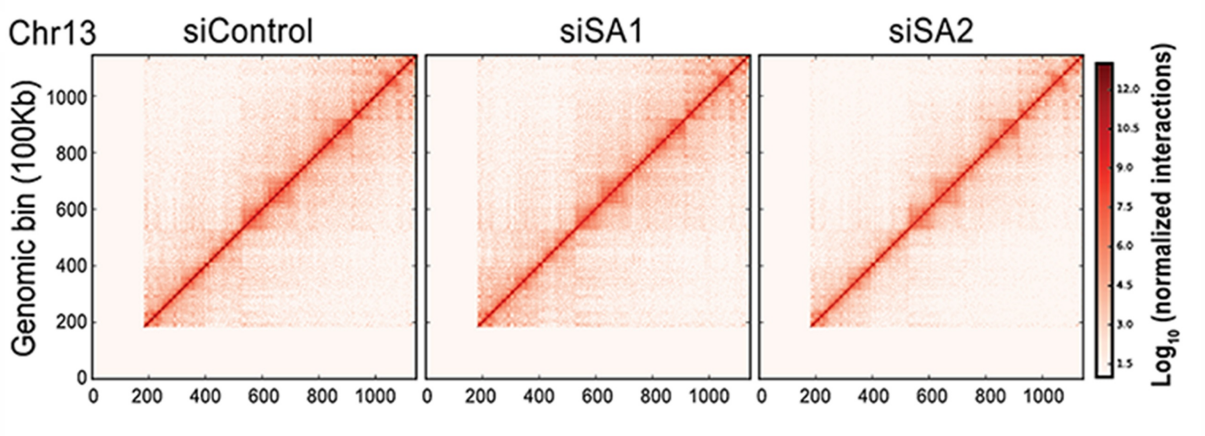


Figure R28. Contact maps for MCF10A cells depleted for cohesin variants

Vanilla normalized Hi-C matrices for chr13 at 100Kb resolution in mock transfected (siC), SA1 depleted (siSA1) and SA2 depleted (siSA2) cells

Depletion of SA1 increased the number of TADs by 204, whereas depletion of SA2 led to a decrease of 439 TADs, when compared to control condition (Figure R29a). Furthermore, we detected minor changes in TAD border strength in both siSA1 and siSA2 conditions, which was slightly more pronounced for the SA1 depletion (Figure 29b). This finding was consistent with our previous results hinting to joint activity of cohesin-SA1 and CTCF at TAD

borders. Interestingly, TAD border conservation was diminished by 25% after SA2 depletion while no significant reduction was caused by depletion of SA1 (Figure R29c). We therefore suggest that some TADs might arise in a CTCF independent manner and instead depend on the interaction of intra-TAD cohesin-SA2 with different transcriptional regulators.

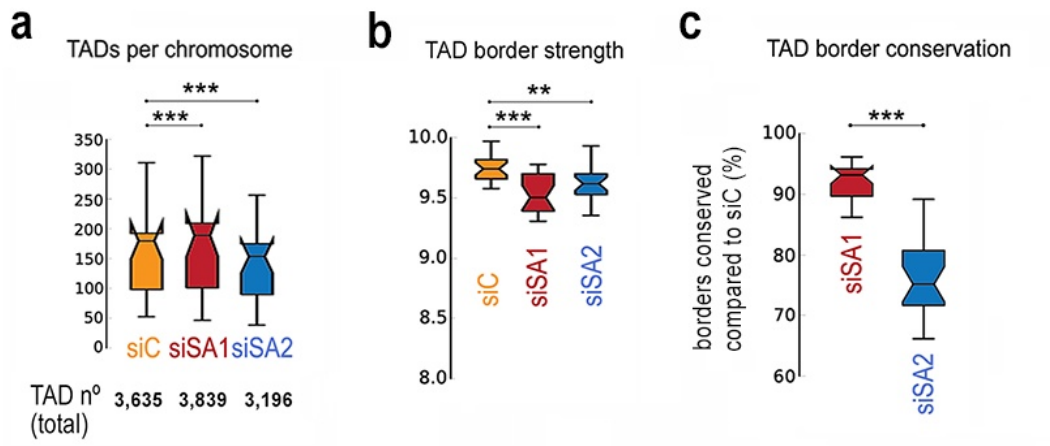


Figure R29. TAD border conservation is reduced when SA2 levels are reduced

Boxplots showing number of TADs per chromosome (a), TAD border strength (b) and TAD border conservation (c). A border was considered conserved between siControl (siC) and siSA1 or siSA2 experiments if it was localized within +/- 2 two bins in both experiments. The box extends from the lower to upper quartile values of the data, with a line at the median. Notches represent the confidence interval around the median.

4.2. Cohesins SA1 and SA2 are involved in different types of chromatin interactions

In view of their different genomic distribution and enrichment in functional elements, we wondered if cohesin-SA1 and cohesin-SA2 could have specific roles in the establishment of distinct types of chromatin interactions involved in different biological processes. Indeed, the analysis of the genomic interactions as a function of the genomic distance evidenced a distinct behaviour for SA1 and SA2 depleted cells compared to control cells. Loss of SA2 increased mid-range contacts (0.1 -1.3 Mb, blue shade in Figure R30) while loss of cohesin-SA1 increased long-range contacts (>1.4 Mb, red shade in Figure R30).

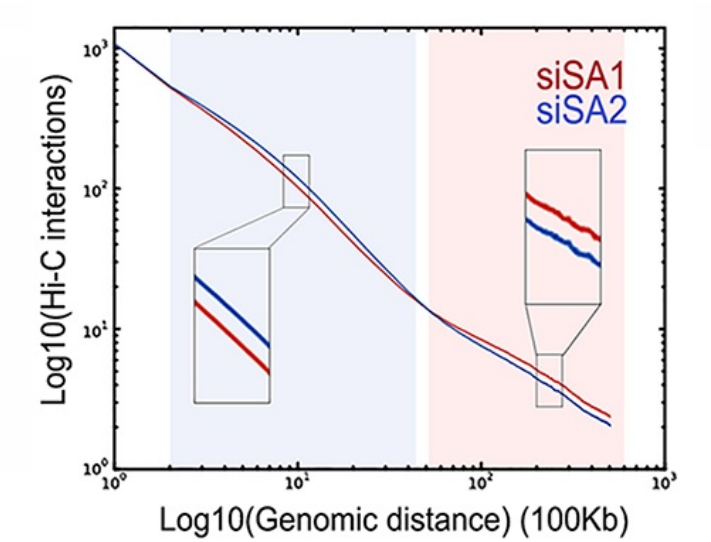


Figure R30. Decay of genomic interactions as a function of distance differs upon downregulation of SA1 or SA2

Interaction decay plot showing Hi-C interactions as a function of genomic distance averaged across the genome for a maximum distance of 50 Mb for siSA1 (red line) and siSA2 (Blue line) treated cells compared to mock treated cells. Shadowed areas highlight intervals in which clear differences are found between conditions.

These distinct effects were also evident in matrices representing gained and lost interactions separately. Gain of long-range interactions for siSA1 condition was observed as signal enrichment far from the diagonal, at distance lengths of several megabases, while interactions decreased close to the matrix diagonal (Figure R31a, upper panels). In contrast, interactions increased in the mid-range (500kb-2Mb range) in the SA2 depleted cells, not so far from the diagonal while loss of interactions occurred at very short distances, very close to the diagonal (Figure R31a, lower panels).

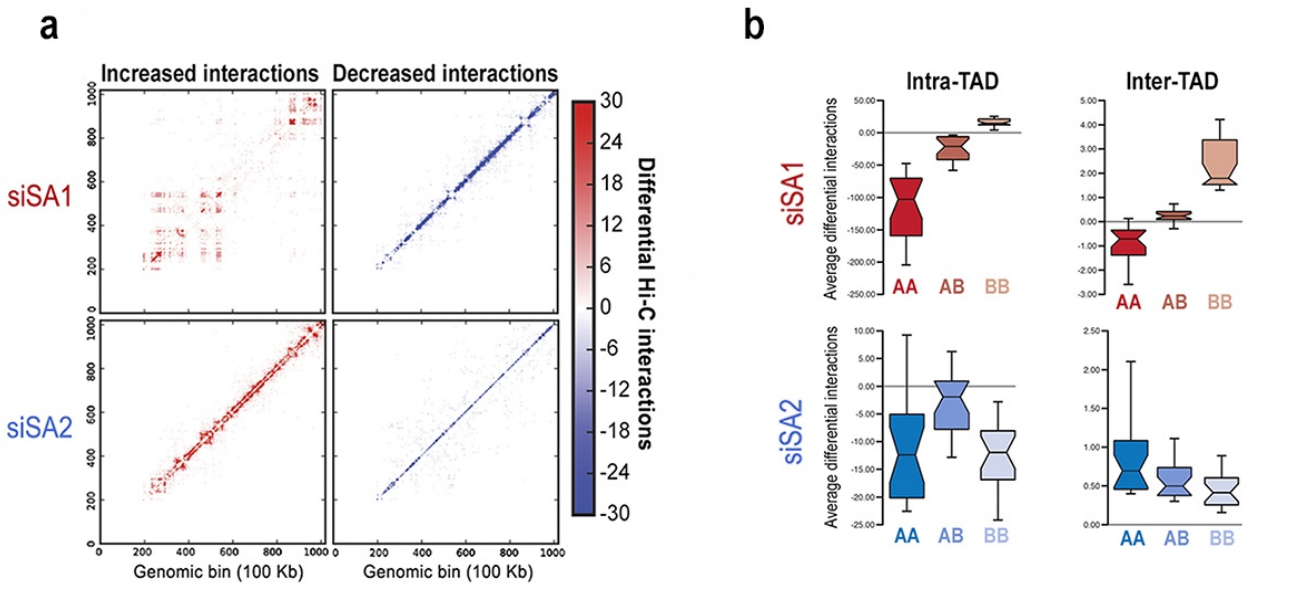


Figure R31. Distinct effect of SA1 and SA2 depletion on long range chromatin contacts

a. Increased (red) and decreased (blue) genome-wide interactions between siSA1 or siSA2 treated cells compared to control cells in chr15.

b. Boxplots showing differential interactions (averaged per chromosome) separating them according to their being inter- or intra-TAD interactions and to their localization in A or B compartment.

When we analysed the localization of lost and gained interactions with respect to A and B compartments, we found that the increased interactions SA1 depleted cells were located within the B compartment, while mid-range interactions were lost in the A compartment (Figure 31b). Moreover, both gained and lost interactions could be inter- or intra-TAD, with no obvious bias. One possible interpretation of these data is that SA1 depletion results in a more "relaxed" A compartment, which is compensated by increased compaction of the B compartment.

In contrast, increased mid-range contacts observed upon SA2 depletion were inter-TAD and mostly within the A compartment, at least in part due to loss of TAD borders. Decreased short-range contacts were intra-TAD, within A and B compartments, that could correspond to enhancer-enhancer or enhancer-promoter interactions, given the prevalence of SA2-only positions in these elements.

In spite of the limitations of the resolution of our Hi-C experiments, and aimed to establish a correlation between cohesin genomic distribution, gene expression and chromosomal organization, we examined the enrichment of several features including TSS, RNA (as a measure of gene expression), cohesin and CTCF positions in A and B compartments, A/B

borders and TAD borders. As expected, all the mentioned features were found to be enriched more than expected at A compartments, with SA2-only positions being more enriched than common and SA1-only positions. At TAD borders, SA1-only and CTCF positions were more enriched than SA2-only positions. Most features were depleted from borders separating A and B compartments as well as from B compartments, except SA1-only positions. This finding suggested that cohesin-SA1 might play a unique role in modulating A/B compartment identity, although, as shown above, no clear compartment switches were found in siSA1 treated cells.

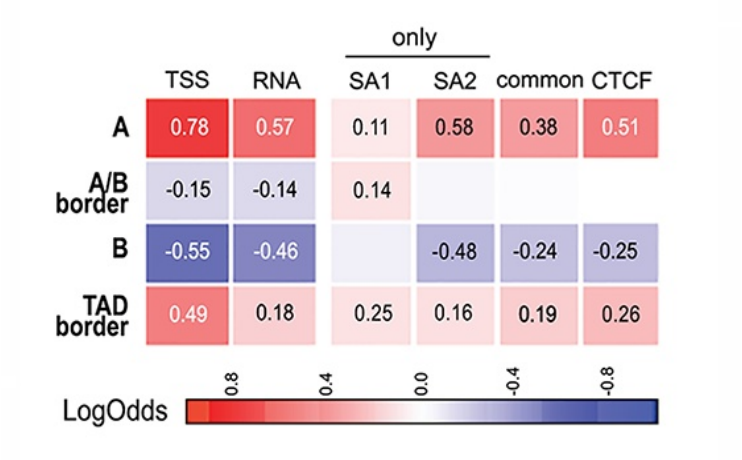


Figure R32. Correlation between cohesin distribution and genome organization

Enrichment of SA1-only, SA2-only and common cohesin sites in AB compartments, A/B borders and TAD borders. CTCF, TSS and RNA (from RNA-seq data) are included as well. Squares with numbers are significant (Fisher exact test, p-values < 0.001).

In summary, we propose that cohesin-SA1 has a more structural role in genome organization, in particular mediating TAD/subTAD formation together with CTCF, while cohesin-SA2 would be more important for functional intra-TAD contacts together with transcriptional regulators. In the absence of cohesin-SA1, cohesin-SA2 can still cooperate with CTCF in genome organization although border strength is decreased and the A compartment is loosened. In the absence of cohesin-SA2, short-range intra-TAD contacts decrease while new contacts are formed between neighbouring TADs and these changes have more noticeable consequences for gene expression.

Discussion

Due to its role in sister chromatid cohesion, cohesin is essential for genome stability and inheritance. In addition, its capacity to hold topologically two DNA segments and bring them into close proximity transcends the mitotic context and is instrumental for mediating folding of the chromatin fibre in interphase cells. The resulting 3D structure is dynamic and contributes to genome function.

All four subunits of cohesin are indispensable for the complex to be functional. Two distinct cohesin complexes that carry one of two isoforms of the SA subunit coexist in somatic vertebrate cells. The fact that these isoforms share high sequence identity could explain why the existence of two cohesin variants has long been neglected in functional studies. However, a few studies have recently pointed to non-redundant activities of the two isoforms in cohesion and embryonic development. Cohesin-SA1and cohesin-SA2 are specifically required for telomere and centromere cohesion, respectively (Remeseiro et al., 2012a; Canudas and Smith, 2009). Importantly, the presence of both cohesin variants is indispensable for proper embryonic development since mouse deficient for SA1 or SA2 are embryonic lethal (Remeseiro et al., 2012a; M. De Koninck and E. Lapi, unpublished results). In this thesis, we have explored the functional specificities of the two cohesin complexes regarding their contribution to 3D chromatin organization and gene regulation. We have observed consistent differences in their genome-wide distribution, especially with respect to colocalization with CTCF, and their enrichment in functional genomic elements. Consistent with this finding, depletion of either isoform has different impact on gene expression, pointing to distinct mechanisms of gene regulation by cohesin-SA1 and cohesin-SA2. We also provide strong evidence that the two complexes contribute differently to the organization of the genome at the megabase scale by mediating different ranges of "*in cis*" genome interactions. While cohesin-SA1 preferentially contributes to the stabilization of TAD boundaries together with CTCF, cohesin-SA2 promotes local contacts between enhancers and promoters together with transcriptional regulators. Importantly, cohesin-SA1 cannot perform this function when SA2 is absent.

1. The different distribution and dynamic behaviour of cohesin-SA1 and cohesin-SA2

We have identified three categories of cohesin positions that are conserved in all the human cell lines that we have examined: common, SA1-only and SA2-only. Results from our group not shown in this Thesis confirm that they are also found in mouse ES cells and MEFs.

1.1 Common cohesin positions

Both cohesin-SA1 and cohesin-SA2 can be found at common positions, which always overlap with CTCF. Although an important fraction of these positions can be found in cis-regulatory elements, a larger fraction occupies a chromatin state defined by the presence of CTCF. These positions have higher cohesin occupancy and are therefore easier to detect.

Since ChIP analyses are performed in a cell population, it is unclear whether at each of these positions there is more than one cohesin complex. It is conceivable that one cell would have cohesin-SA1 while its neighbour would have cohesin-SA2 at a given position at the time of fixation, resulting in detection of the two complexes at that genomic location. However, our Re-ChIP analyses indicate that, at least in several common positions, more than one complex is indeed present. One can imagine two configurations for this situation, one with two complexes co-entrapping the two segments at the base of a loop and another with two interacting complexes each embracing a segment ("stacking" and "handcuffed", respectively, in Figure D1a). As described in the Introduction, it is likely that cohesin arrives at these CTCF-bound sites through loop extrusion after being loaded elsewhere (Busslinger et al., 2017). The presence of CTCF proteins could represent an obstacle for cohesin progression, either physical, as suggested by in vitro data (Stigler et al., 2016; Davidson et al., 2016), or functional. Recent data support the idea that Pds5 is required for cohesin to arrest at CTCF sites (Wutz et al., 2017) and it has been suggested that these sites would favour the replacement of NIPBL -which is required for loop extrusion- by Pds5 (Petela et al., 2017; Haarhuis et al., 2017).

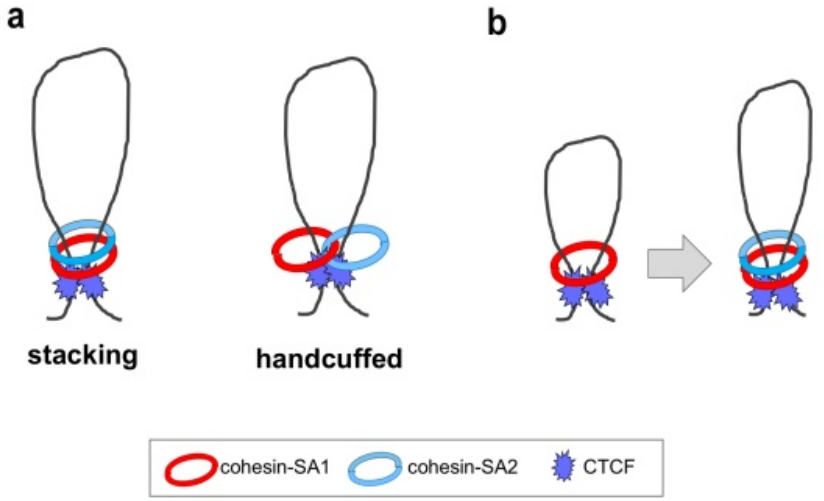


Figure D1. Cohesin configuration at common positions

See text for details.

Single-molecule imaging approaches have reckoned that CTCF binds chromatin much more dynamically than cohesin, ~1 min vs. ~22 min residence time (Hansen et al., 2018). Indeed, binding of CTCF and cohesin to chromatin is very different. CTCF establishes non covalent interactions with specific DNA sequences through its zinc finger domains (Hashimoto et al., 2017) while cohesin has to be loaded topologically (Haering et al., 2008) and it is still unclear where the chromatin fiber is located inside the ring and what kind of interactions the different subunits establish with the DNA. In any case, it seems likely that once a loop has been formed, cohesin stays in place even if CTCF dissociates and rebinds, at least for several minutes. The sharp and narrow profiles of average density plots at common sites agree with the idea of CTCF being the major determinant of cohesin positioning. Both SA1 and SA2 show the same behavior at these sites, suggesting that they both are stopped by CTCF in a similar way. Alternatively, cohesin-SA1 may be more easily stopped by CTCF and, once there, it may retain cohesin-SA2 complexes arriving afterwards (Figure D1b). This idea is in fact consistent with the observed redistribution of cohesin-SA2 in SA1-deficient MEFs, which results in detection of two times more cohesin binding sites compared to wild type MEFs. The new sites are featured by low cohesin occupancy and lack of CTCF colocalization (Remeseiro et al., 2012b). Thus, cohesin-SA2 can still occupy common positions on its own when SA1 is depleted, but may be stopped at these positions for shorter times. To test this hypothesis, it would be interesting to repeat the imaging experiments described above (Hansen et al., 2018) for SA1 and SA2-containing complexes in wild type cells and cells lacking either SA1 or SA2.

Why would be cohesin-SA1 retained preferentially at CTCF sites? We can entertain several possibilities. For instance, in vitro studies showing recombinant cohesin complexes moving along naked DNA, mentioned above, suggest that DNA-bound cohesin adopts a conformation in which the size of the pore is relatively small, thus hindering its ability to bypass DNA-bound proteins such as CTCF (Stigler et al., 2016, Davidson et al., 2016). Similar studies comparing the two cohesin variants could elucidate whether the two cohesin variants assume different conformations when bound to DNA and determine their ability to bypass different obstacles. Alternatively, if cohesin's arrest at CTCF sites depends on the replacement of NIPBL by Pds5, the switch could be easier for cohesin-SA1 than cohesin-SA2. While both SA1 and SA2 containing complexes can associate with cohesin regulators (Pds5A/B, WAPL, Sororin), immunoprecipitation analyses suggest that the affinity for these interactors could be different (Kim et al., 2016). In our study, for instance, SA2 immunoprecipitates pull down more WAPL.

The two variants could also have different affinities for CTCF. Only one study has dissected the interaction between CTCF and cohesin biochemically (Xiao et al., 2011). In vitro pull down experiments indicated that a fragment of SA2 comprising amino acids 162 to 290 could interact with CTCF. Interestingly, residues next to this region (K290, D326 and K330 are key for binding to WAPL, cohesin unloading factor (Hara et al., 2014). Thus, there could be some competitive binding of WAPL and CTCF to the SA subunit, and the balance could be slightly different for SA1 and SA2, even though both proteins are highly homologous along these regions. Our data indicate that the ratio between cohesin and WAPL is much lower in common than in SA2-only positions, again suggesting that the presence of CTCF impairs WAPL accessibility and as a consequence, stabilizes cohesin binding at CTCF positions (Figure D2).

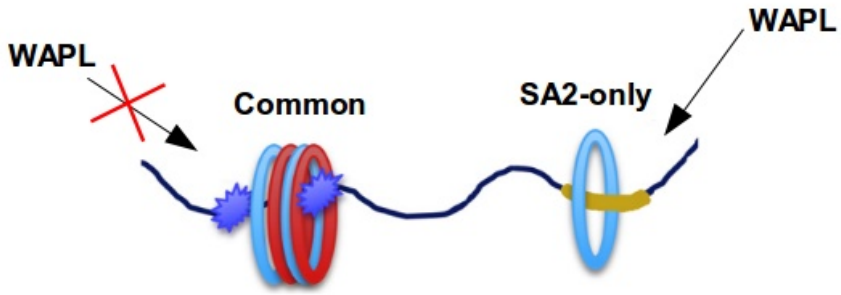


Figure D2. Different determinants of cohesin binding to chromatin

Binding of cohesin at common positions is determined by CTCF while binding of cohesin at SA2-only positions may depend on the presence of transcriptional regulators located at enhancers. The presence of CTCF or the observed stacking of cohesin rings at common positions could restrict WAPL accessibility.

The calibrated ChIP-seq experiment that we performed in MCF10A cells depleted of SA1 or SA2 has also shown that removal of any of the cohesin variants leads to increased accumulation of the other variant at common positions. This could reflect the attempt to compensate and ameliorate possible structural problems related with the function of cohesin complex at those sites. Intriguingly, we did not observe changes in chromatin bound levels of cohesin variants under those conditions, as measured by immunoblot. Thus, we speculate that the observed increase could be due to the redistribution of the remaining cohesin variant from “background” locations. It is also unclear whether this increase fully restores the levels of cohesin in control condition – a calibrated ChIP-seq analyses of a common cohesin subunit such as SMC1 should be performed to answer this question- or if both variants have the same ability to form TAD boundaries together with CTCF in common sites. As discussed

in the last section, this may not be the case, since SA1 depletion affects TAD border strength.

1.2 SA2-only cohesin positions

SA2-only positions are clearly different from common positions. They lack SA1 and CTCF, are mainly located in enhancers, have lower cohesin occupancy and are more difficult to detect. The existence of CTCF-independent cohesin positions had been previously described in human and murine cells (Schmidt et al., 2010; Kagey et al., 2010). Our work demonstrates that cohesin-SA2 is the variant present at these sites. Schmidt et al. (2010) identified a fraction of cohesin in human breast and liver cancer cells that did not localize with CTCF but with the estrogen receptor (ER) and liver-specific transcription factors, respectively. They concluded that “cohesin co-binds across the genome with transcription factors independently of CTCF (…) and may help to mediate tissue-specific transcriptional responses via long-range chromosomal interactions”. Later on, the same group mapped cohesin, CTCF, and a collection of tissue-specific and ubiquitous transcriptional regulators using ChIP-seq in primary mouse liver and confirmed that cohesin sites lacking CTCF correlated instead with the binding of master regulators and enhancer marks and were significantly associated with liver-specific expressed genes (Faure et al., 2012). Interestingly, three different cohesin subunits were immuno-precipitated from chromatin, Rad21, SA1 and SA2. Although the authors decided to define cohesin presence as the occurrence of at least one of its subunits, the prevalence of SA2 over SA1 in these so called “cis-regulatory” modules was clear (Figure D3).

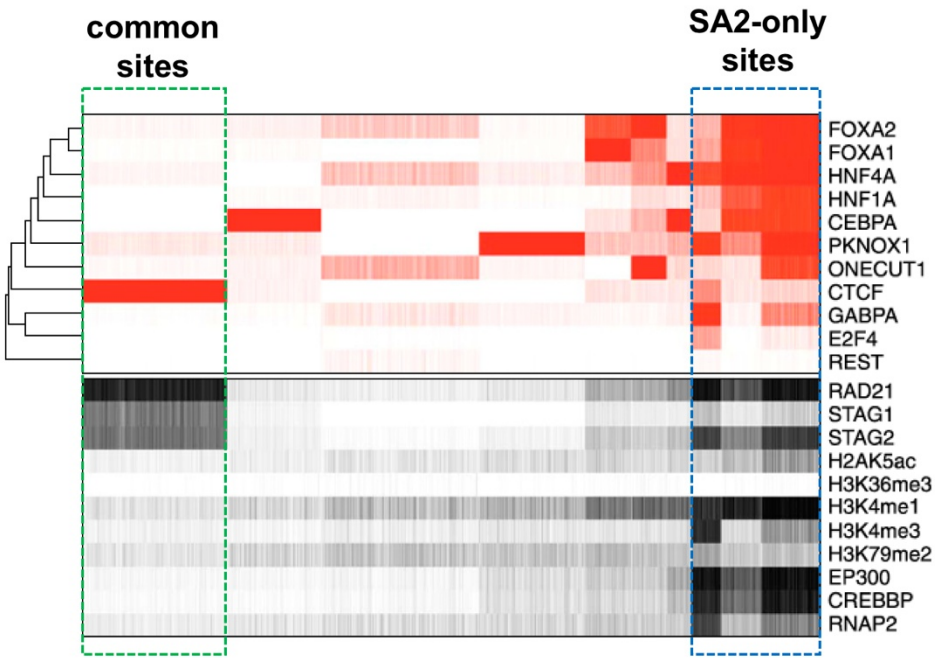


Figure D3. Cohesin-SA2 is the variant present at cohesin-non CTCF sites

Results from the study by Faure et al. (2012). Top, Results from K-means clustering (K = 10) of the binary presence/absence of ChIP-seq peaks of the indicated transcription factors within cis-regulatory modules (CRMs) in mouse liver cells. The clusters were indexed and sorted by the proportion of CRMs with cohesin-non-CTCF peaks in each cluster (increasing from left to right). Bottom, the binary presence/absence of ChIP-seq peaks for cohesin and the indicated chromatin features visualized according to the K-means results as above. We have overlay squares showing common and SA2-only cohesin positions.

Average read density plots of SA2-only positions are broader, which could indicate that the way cohesin arrives and stops at these positions is different from the one described above for common positions. Kagey et al. (2010) first showed the presence of cohesin connecting the enhancers and promoters of pluripotency genes in mouse ES cells together with Mediator in order to maintain active transcription. NIPBL was found at the same sites and was shown to interact physically with cohesin and Mediator, suggesting that cohesin loading could occur directly at these sites. In such scenario, one could envision cohesin stabilizing an enhancer-promoter loop already established by Mediator, transcription factors and RNA polymerase II. The exact position of the loaded cohesin complex could be less fixed and thereby give rise to the broader average profiles observed (Figure D4).

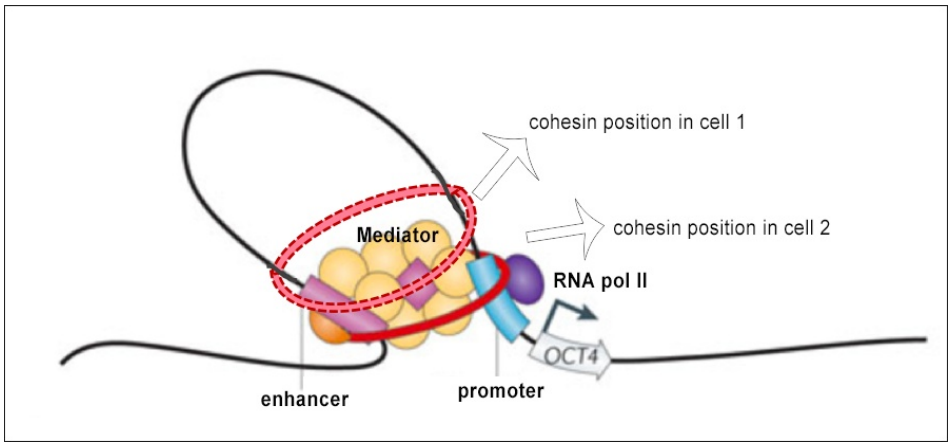


Figure D4. The exact position of cohesin-SA2 at enhancer-promoter loops is variable

Unlike common positions, in which the binding of CTCF fixes the positions of the coheisn complex, the precise position of cohesin-SA2 at enhancer promoter loops could be slightly different in cells in the population. We suspect that arrest of cohesin could also be shorter.

Why is cohesin-SA2 the preferred variant in these local chromatin loops involving transcriptional regulators? One possibility is that, as discussed above, cohesin-SA1 becomes more easily retained at CTCF positions leaving less soluble cohesin-SA1 complex available for NIPBL to load at these sites. Another is that NIPBL may have higher affinity for cohesin-SA2. In vitro, the SA subunit is essential for cohesin loading although its function in the process remains unclear (Murayama et al., 2014). Recent studies report that SA1 and SA2 can bind double stranded (ds) and single stranded (ss) DNA, with a preference for the latter that is more clear in the case of SA2 (Lin et al., 2016; Countryman et al., 2018). In one of these studies, the SA2 protein was shown to switch between a diffusion mode along dsDNA to stable binding at ssDNA gap. Whether SA1 can do the same was not investigated. Interestingly, an in vitro reaction that recapitulates establishment of DNA-DNA interactions by a cohesin ring already embracing dsDNA has shown that the second DNA molecule must be ssDNA (Murayama et al., 2018). Taken all these evidences together, we speculate that initial binding of cohesin to DNA before topological entrapment likely depends on the SA subunit and each variant may display different specificities regarding secondary structures of the DNA or the presence of single stranded nucleic acids (DNA or even RNA) resulting from active transcription. Finally, protein-protein interactions with transcriptional regulators present at enhancers could also drive preferential loading of cohesin-SA2 at these sites. The homology between SA1 and SA2 decreases in the N- and C-terminal regions (see Figure I10 of Introduction), and such regions could be responsible for binding specific interactors. Indeed, N-terminus of SA1 interacts with telomeric protein TRF1, which enables cohesin-

SA1 to mediate telomeric cohesion (Canudas et al., 2007; Canudas and Smith, 2009). Similarly, SA2 may interact with certain transcriptional regulators through its unique regions.

1.3 SA1-only cohesin positions

SA1-only sites have significantly higher enrichment of SA1 than SA2 and co-localize with CTCF. At least in HMEC cells, almost half of the SA1-only sites are located in chromatin states corresponding to “heterochromatin/low signal” regions that could correspond to regions with low accessibility to the antibodies used for ChIP-seq. The presence of some SA2 and CTCF makes us think that these could be in fact a subset of common positions in which detection of SA2 is less efficient than detection of SA1 (Figure D5). However, we cannot rule out that SA1-only sites could be specialized cohesin positions with a unique structural function, different from that of common positions. Supporting this last hypothesis is the finding that SA1-only regions are enriched in the transitions between A and B compartments (Figure R32). Also, unlike common positions, cohesin-SA1 sites are not conserved between cell types. What is clear from our results is that the function of cohesin-SA1 is tightly linked to CTCF.

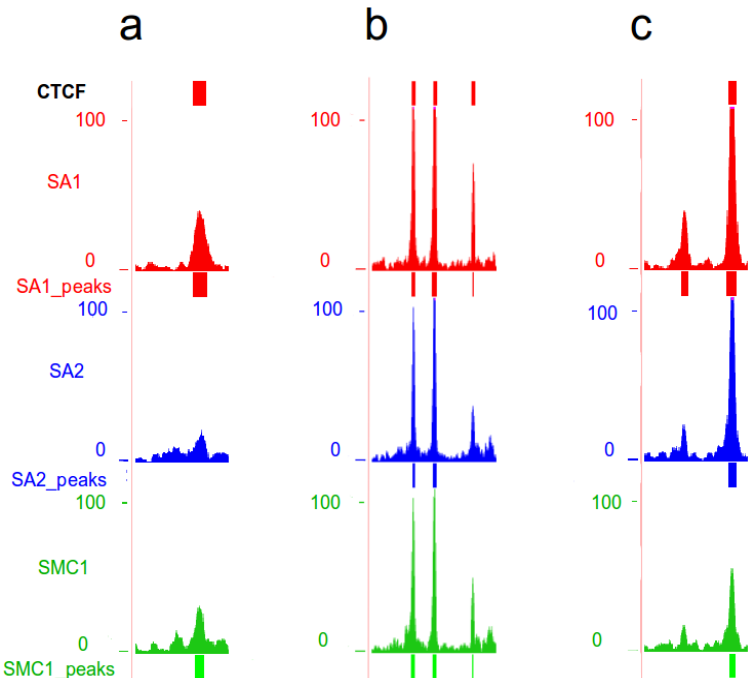


Figure D5. SA1-only cohesin positions

Snapshot of the genome browser showing examples of SA1-only positions. Distribution of CTCF, SA1, SA2 and SMC1 from HMEC is shown. SA1-only positions are shown next to common cohesin positions in **b** and **c**.

2. Cohesin-SA2 regulates cell-type specific gene expression

The analysis of cohesin variants enrichment in different chromatin states has strengthened our hypothesis about their different functions. The high association of SA2-only positions with enhancers, in particular superenhancers, and with gene promoters, has corroborated previous findings showing enrichment of CTCF-less cohesin positions in regulatory elements (Kagey et al., 2010; Schmidt et al., 2010; Faure et al., 2012; Figure D3). On the other hand, the observed association of common and SA1-only positions with insulators was expected, given the almost complete overlap of these positions with CTCF. It is important to highlight that SA1-only positions, even when their number differed between the cell lines in the study, behaved similar to common positions, with the exception of their enrichment in heterochromatic/poorly defined regions, whose relevance is still an open question as discussed above.

We were interested in identifying potential SA2-specific DNA-binding factors that would account for its particular genomic distribution. Both our motif enrichment and proteomic analysis strongly indicate that cohesin-SA2 could fulfil gene regulatory functions in a CTCF-independent manner through its association with different factors. Moreover, transcriptional regulators such as ZMYM2 and YAP1 were found to interact specifically with cohesin-SA2 and we showed that they bound at cohesin SA2-only sites. One important limitation of the immunoprecipitations analysis is the difficulty of preserving cohesin interactions that occur in the context of 3D folded chromatin. The use of detergents, nucleases or sonication to release cohesin from chromatin will necessary disrupt chromatin topology and therefore impede protein-protein interactions that occur in a particular chromatin conformation. In fact, we were not able to get CTCF in our proteomic analysis. Similarly, we did not find ZMYM2 and YAP1 binding motifs in the motif enrichment analysis of SA2-only positions. In this regard, it is worth mentioning a study proposing that cohesin would stabilize the binding of transcription factors to lower-affinity sequence motifs (Faure et al., 2012). In addition, the ways in which cohesin interacts with transcription factors are still largely unknown.

Recent work from our group has supported the hypothesis that CTCF-less cohesin positions are involved in cell-type specific gene regulation (Cuadrado et al., 2015). In this study, the genome-wide distribution of cohesin in two murine tissues, brain and pancreas, was explored and correlated with their tissue-specific expression programmes. Several major conclusions from this study were important for this thesis: 1) CTCF-bound cohesin sites are more invariant between tissues than cohesin non-CTCF sites; 2) tissue-specific cohesin positions display better overlap with marks of active enhancers and promoters (H3K4Me1

and H3K4Me3) than common cohesin sites 3) tissue-specific cohesin is more frequently found at active tissue-specific genes than tissue-specific CTCF (Figure D6). In that study, ChIP-seq was performed using SMC1 and SA1 antibodies, but no SA2, so we can only indirectly infer which cohesin positions correspond to each variant. The fact that that many active CTCF-less promoters contain SMC1 with no or little SA1 support the existence of SA2-only positions at those active promoters. On the other hand, active promoters that were bound by both cohesin and CTCF displayed similar frequency of both SMC1 and SA1.

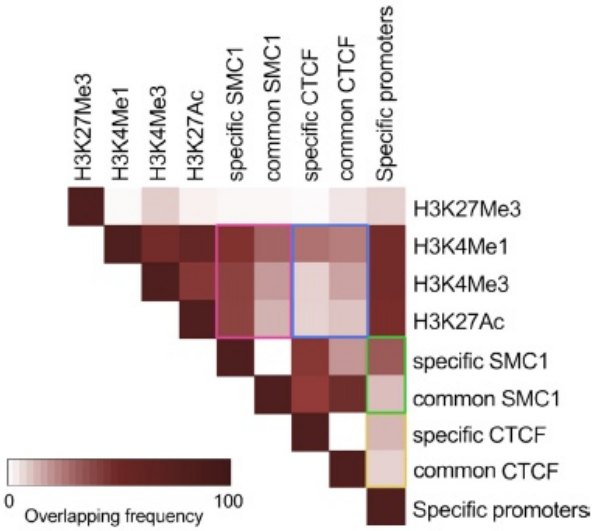


Figure D6. Cohesin distribution correlates with tissue-specific transcription

Heat-map visualization of the pairwise overlap between cerebral cortex chromatin marks and cortex-specific and non-specific ('common') SMC1 and CTCF positions. Cortex-specific cohesin (SMC1) binding sites display a better overlap with enhancer and promoter marks (H3K4Me1 and H3K4Me3) than ubiquitous cohesin binding sites (purple rectangle). This is not the case when comparing the overlap of cortex-specific and common CTCF positions with those same chromatin marks (blue rectangle). Moreover, cortex-specific cohesin is more frequently found at the promoters of cortex specific genes than cortex-specific CTCF (compare green and yellow rectangles). Taken from Cuadrado et al., 2015.

Based on these previous results, we explored the cell-type specificity of cohesin variants distribution between the two human primary cell lines from different embryonic origin used in our study, HMEC and HCAEC. While common positions were highly conserved, SA2-only positions were mostly cell-type specific. This result, together with the enrichment of SA2-only sites in superenhancers points to a critical role of cohesin SA2 in the control of the transcriptional programs that define cell identity (Hnisz et al., 2013). It was surprising to find that, in spite of showing clear similarities to common positions, SA1-only positions are cell-

type specific, which reinforces the above discussed possibility that these positions could have a different and still unexplored role.

Transcriptome analyses after depletion of the two cohesin variants have shown that reduction of cohesin-SA2 levels alters the expression of about 700 genes in MCF10A cells, while depletion of cohesin-SA1 affects the expression of a little over 150 genes. The small overlap between the genes whose transcription changes in SA2 and CTCF depleted cells further confirmed a CTCF-independent function of cohesin-SA2. In hematopoietic precursors, knock down of SMC1 or SA2 led to similar gene deregulation driving premature commitment to the myeloid lineage (Mullenders et al., 2015), consistent with a more clear role of cohesin-SA2 in gene regulation. On the other hand, the modest effect of SA1 depletion on gene expression may be due to the fact that, under this condition, the function of cohesin-CTCF exerted common positions can be performed by cohesin-SA2. This function is likely related with the maintenance of TAD boundaries and might have little impact on overall transcription. In fact, even when most cohesin is removed, the consequences on gene regulation are not that compelling. For instance, genetic deletion of the cohesin loader NIPBL from mouse liver cells reduced significantly cohesin levels and disrupted TAD integrity but led to modest changes of gene expression affecting around 1,000 genes (Schwarzer et al., 2017). The authors showed that observed changes were likely due to impaired communication of regulatory elements with target genes, even though the regulatory capacity of the enhancers was preserved. Acute removal of cohesin in human HCT116 cells using a degron system combined with measurement of nascent transcription by PRO-seq revealed that only genes linked to superenhancers were strongly downregulated whereas most other genes were mildly affected (Rao et al., 2017). Although these results seem to argue against a major role of cohesin in facilitating enhancer-promoter communication, we should bear in mind that small changes in transcription of key genes may be physiologically relevant in specific cell contexts, such as a cancer cell (Flavahan et al., 2016; Shen et al., 2016), or when the cell needs to respond to developmental cues, hormone signalling or other environmental stimuli (Stadhouders et al., 2018; Le Dily et al., 2014; Antony et al., 2015; Xu et al., 2016). Zuin et al. (2014) first showed that depletion of CTCF or depletion of cohesin (RAD21) altered the expression of distinct groups of genes. We found that only one third of the genes deregulated in siSA2 treated cells were also deregulated upon depletion of CTCF in MCF10A cells. Reduction of CTCF levels caused deregulation of a larger number of genes (1191 in siCTCF vs 716 in siSA2 treated cells). A recent study reported changes in around 5,000 genes upon acute, near complete removal of CTCF by a degron system in mouse ES cells (Nora et al., 2017). However, it is important to note that the number of genes

showing altered expression 1 day after inducing cohesin degradation were only 370, 13-fold less than at day 4. Thus, secondary effects can rapidly become confounding making it difficult to asign the cause of a transcriptional change to a change in chromatin contacts. The use of degrons has proven in the last years to be a very efficient and reversible system to eliminate a protein in an acute way (Nishimura et al., 2009; Natsume et al., 2016; Nora et al., 2017; Rao et al., 2017; Wutz et al., 2017). This methodology overcomes the problems related with incomplete depletion obtained by other techniques such as siRNA transfection and are particularly suited for stable proteins like cohesin. We plan to develop degron systems to deplete each of the cohesin variants in the near future, in order to further improve the study of SA1 and SA2 specific functions.

Apart from the magnitude of the transcriptional changes, we were interested in their entity. The significant up-regulation of pathways specific for hematopoietic and nervous system together with the aberrant expression of several important cell-type specific transcription factors observed in MCF10A cells with reduced levels of SA2 (e.g., of Irx3 and Tfap2c or BDNF) agrees with the role of cohesin-SA2 in maintaining cell identity (Figure D7).

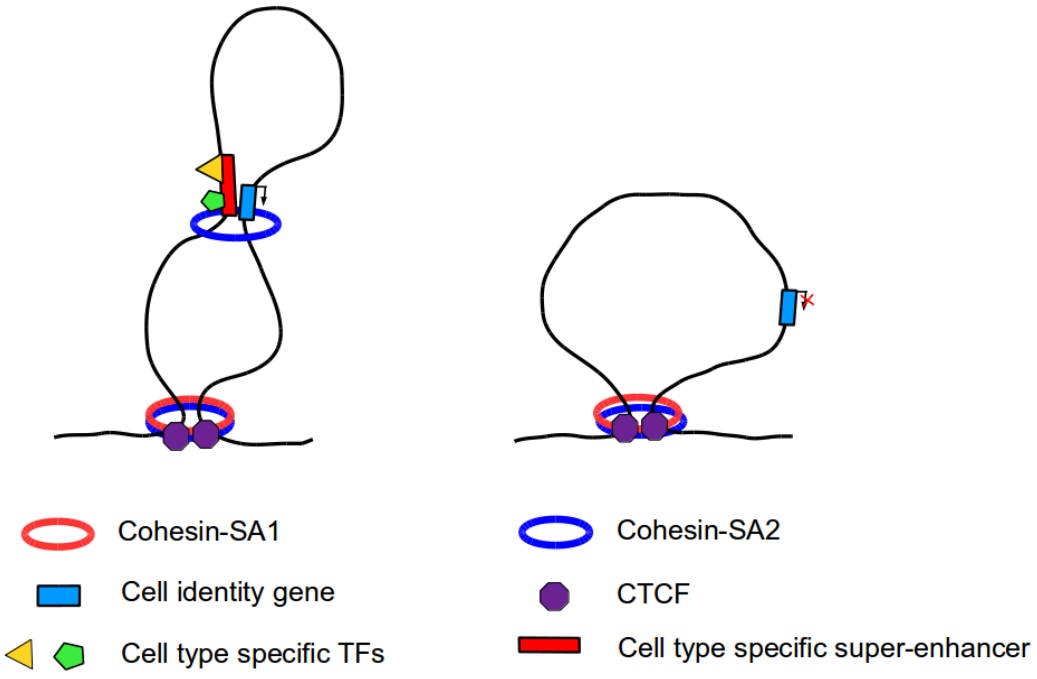


Figure D7. Cohesin-SA2 is responsible for maintaining cell identity

SA2-only cohesin positions are largely cell-type specific and enriched in cell-type specific super-enhancers. Cohesin-SA2 regulates expression of cell-identity genes from SA2-only positions in a CTCF-independent manner, through interaction with different TFs (left). Absence of cohesin-SA2 leads to dysregulation of cell-identity genes (right).

Such role is also consistent with the hypothesis that loss of function cohesin mutations identified in myeloid malignancies are associated with skewing of expression programs towards precursor cells and their impaired differentiation potential, as mentioned above (Mazumdar et al., 2015; Mullenders et al., 2015; Viny et al., 2015). It has also been suggested that cohesin works together with Polycomb-group proteins in these hematopoietic precursors to repress *Hox* genes (Fisher et al., 2017; Li et al., 2017). Unpublished results from our group indicate that cohesin-SA2 is the major cohesin variant present at Polycomb-bound regions in mouse ES cells. These results support the promiscuous role of SA2 in the control of gene expression, being involved both in activation and repression of gene transcription, likely through its ability to associate with transcriptional regulators.

Finally, by analysing the S100 gene cluster in different cell lines, we have observed a strong correlation between the presence of SA2-only positions, the intensity of local interactions and the levels of expression of nearby genes. Moreover, several genes of the cluster that contain cohesin-SA2 only positions at their promoters in MCF10A cells become downregulated after reduction of SA2 levels. This findings support the idea that the presence of cohesin-SA2 plays an important role for the establishment and/or maintenance of the physiological regulatory contacts between genes and their corresponding regulatory elements.

3. Cohesin variants make different contributions to genome organization

The emergence of 3C techniques, especially Hi-C, revolutionized our understanding of genome organization (Dekker et al. 2002; Lieberman-Aiden et al. 2009; Dixon 2012; Nora 2012). Current opinion in the field holds that cohesin-mediated loop extrusion within defined genomic regions is a plausible mechanism of genome organization at the megabase scale (Sanborn et al., 2015; Fudenberg et al., 2016). Consistent with this idea, TADs and prominent Hi-C loops largely disappear when cohesin or CTCF are strongly depleted (Gassler et al., 2017; Nora et al., 2017; Rao et al., 2017; Wutz et al., 2017; Schwarzer et al., 2017).

Efficient removal of CTCF leads to disappearance of TADs and loops formed between two CTCF sites since cohesin no longer encounters loop delimiting factors (Figure D8, top left). However, overall scaling of Hi-C contact frequency as a function of genomic distance is not disturbed, meaning that CTCF does not affect general chromatin compaction (Nora et al., 2017; Wutz et al., 2017; Figure D8, top right). On the other hand, efficient removal of cohesin

from chromatin leads to diminishment of TADs and loops accompanied by a significant reduction of contact probabilities at scale of 10-100s of kilobases that corresponds compaction of genome at a megabase scale (Schwarzer et al., 2017; Figure D8, bottom).

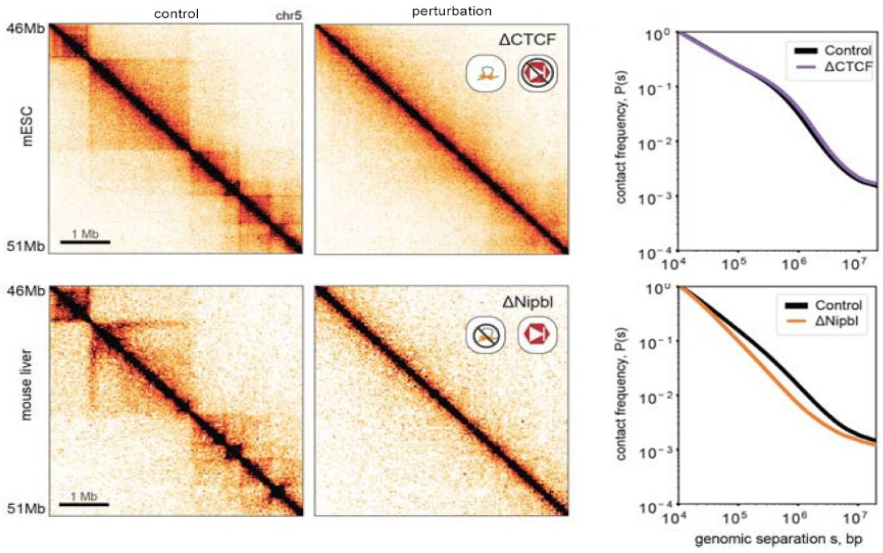


Figure D8. Cohesin and CTCF are required for loop and TAD formation

Hi-C maps for the indicated region in chromosome 5 in mouse ES and liver cells before and after depletion of CTCF (top) and the cohesin loader NIPBL (bottom), replotted from Nora et al, 2017, in which an auxin-inducible degron system was used to deplete CTCF in mESCs and Schwarzer et al. 2017 who used CRE-inducible gene deletion in mouse liver cells to deplete Nipbl and thereby remove cohesin from chromatin. Right, P(s) for the indicated perturbation compared to unperturbed P(s) normalized to contact frequency at 10 kb. Adapted from by Fudenberg et al., 2017.

Even though overall compartmentalization is still present upon removal of CTCF and cohesin from chromatin, cohesin removal leads to enhanced and finer compartmentalization (Nora et al., 2017; Rao et al., 2017; Schwarzer et al., 2017; Wutz et al., 2017). This enhanced compartmentalization is reflected in the emergence of smaller, B-like regions within bigger A-compartments. Additionally, these regions show clear tendency to interact with other regions of the same identity. Interestingly, compartmentalization profile of cohesin-depleted Hi-C maps correlated better with the local activity and chromatin state than the profile of wild type Hi-C maps (Schwarzer et al., 2017). Conversely, compartmentalization is less pronounced in cells lacking cohesin unloading factor WAPL (Gassler et al., 2017; Haarhuis et al., 2017; Wutz et al., 2017). Thus, the intrinsic tendency of chromatin to self-organize based on the local epigenetic landscape and transcriptional activity is likely counteracted by cohesin loop extrusion activity, which brings together loci with different histone modification patterns. Contrary to the initial idea of compartmentalization being a higher hierarchy over TADs in the

genome folding process, it now seems that they are two distinct modes of genome organization that arise through different mechanisms.

By performing Hi-C analyses in SA1- and SA2-depleted cells, we have demonstrated for the first time that the two cohesin variants make different contributions to genome organization. The changes that we observed upon downregulation of one or the other cohesin variant are, not surprisingly, different from those observed after removal of all cohesin, discussed in the previous paragraph. The amount of total cohesin left on chromatin in siSA1 and siSA2 treated cells is very similar, making unlikely that these changes can be explained by the abundance of the complex. We have found that the identity of compartments is largely unperturbed in both conditions, as expected from previous results showing that near complete removal of cohesin is necessary to cause mild switches in compartment identity. TAD border strength showed some decrease upon removal of cohesin-SA1, maybe because the remaining cohesin-SA2 is sufficient to establish TAD borders together CTCF but does not perform this function as well as cohesin-SA1.

Interestingly, TAD-border conservation was significantly reduced upon removal of cohesin-SA2. A recent study has shown that almost 20% of TAD borders are maintained after elimination of CTCF implying that they can be established by some CTCF-independent mechanism (Nora et al., 2017). Moreover, the highest resolution Hi-C maps reported to date have revealed the existence of three classes of TAD boundaries: (1) bound by CTCF, (2) lacking CTCF and proximal to active promoters, and (3) without CTCF or active marks, present in repeat containing regions. Cohesin is present in the first two categories (Bonev et al., 2017). Having in mind these results and the results shown in this thesis, we can speculate that some TAD borders might depend on the interaction of cohesin-SA2 with different transcriptional regulators. Thus, removal of cohesin-SA2 would lead to their disappearance. Unfortunately, the 40-kb resolution of our Hi-C analyses was insufficient to test this hypothesis, since we do not have an accurate map of the TAD borders to correlate with cohesin-SA2-only positions.

Analysis of the genomic interactions as a function of genomic distance together with inspection of Hi-C matrices displaying differential interaction between each depletion and the control cells confirmed that the roles of cohesin-SA1 and cohesin-SA2 in genome folding are not equivalent. Removal of cohesin-SA1 caused an increase in long-range interactions (> 1.4Mb) and a decrease in short- and mid-range interactions (Figure R31). Interestingly, gained interactions occurred in the B compartment while lost interactions were in the A compartment. One possible interpretation of these data is that SA1 depletion results in a

more ‘relaxed’ A compartment, which is compensated by increased compaction of the B compartment (Figure D9). The observed situation could be attributed to reduced integrity of TAD borders when they rely on cohesin-SA2. On the other hand, removal of cohesin-SA2 caused an increase in mid-range contacts (0.1 -1.3 Mb) and a decrease in short-range contacts, very close to the diagonal of the Hi-C matrices (Figure R31). It is likely that lost intra-TAD interactions correspond to enhancer-enhancer and enhancer-promoter contacts mediated by cohesin-SA2. In contrast, the observed increase in inter-TAD interactions, mostly within A compartment, could be in part a consequence of reduced conservation of non-CTCF TAD borders, and consequential emergence of new, inter-TAD interactions (Figure D9). A higher resolution Hi-C experiment would allow us to test dome f these ideas. We have tried to analyse anchors of lost and gained interactions, but with our resolution we were not able to pinpoint the correct cohesin positions involved in the interactions. In line with this, a very recent study has shown that in the absence of cohesin-mediated loop extrusion new chromatin loops emerge connecting regulatory elements over long distances (Vian et al., 2018). Interestingly, only half of these loops have CTCF at both anchors. Another explanation comes from the fact that single-cell Hi-C experiments have demonstrated a high level of stochasticity in TAD border definition among cells in the population. These effects are then averaged into population Hi-C maps. Upon depletion of cohesin-SA2, local contacts could be more disparate among cells in the population making more difficult TAD definition and thereby decreasing the number of TADs and TAD border conservation.

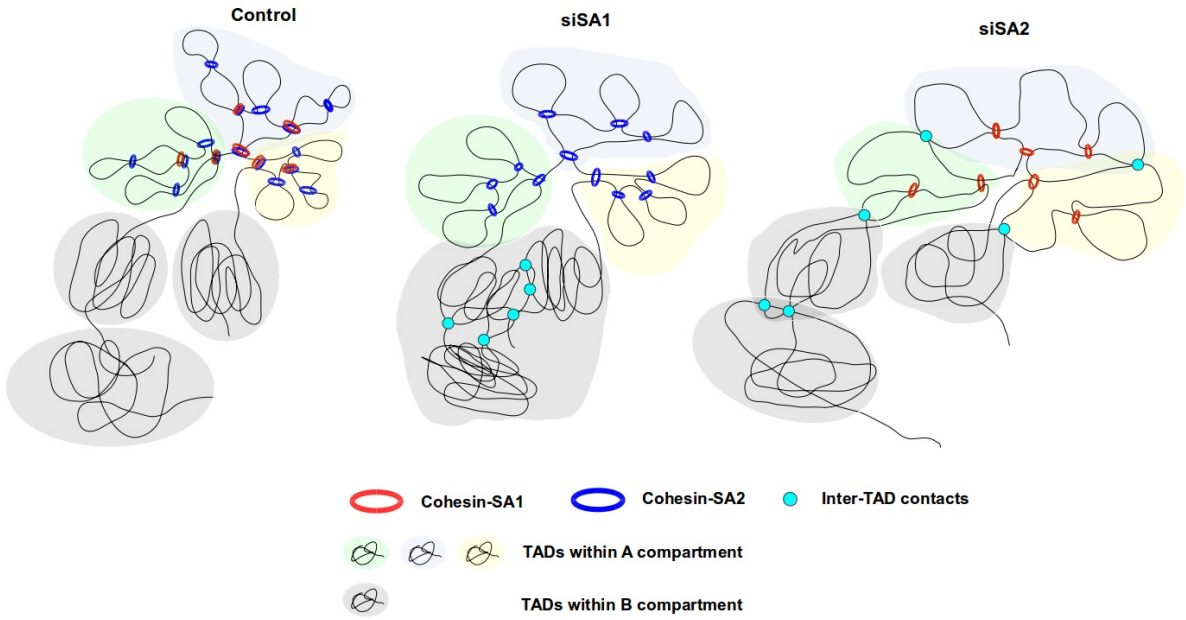


Figure D9. Cohesin variants make different contributions to genome organization

Simplified model of the role of the coheisn-SA1 and cohesin-SA2 in genome organization. Cohesin-SA1 mediatesTAD/subTAD formation together with CTCF, while cohesin-SA2 would be more important for functional intra-TAD contacts together with transcriptional regulators. In the absence of cohesin-SA1, cohesin-SA2 can still cooperate with CTCF in genome organization although border strength is decreased and the A compartment is loosened while the increased long-range contacts are detected in the B compartment. In the absence of cohesin-SA2, short-range intra-TAD contacts decrease while new contacts are formed between neighbouring TADs.

4. Cohesin variants in human disease

The work presented in this thesis brings us closer to understanding pathological consequences of cohesin mutations identified in different cancer types. Given the presence of cohesin-SA2 at regulatory elements and in particular at super enhancers and the inability of cohesin-SA1 to occupy such positions, we propose that tumour cells profit from deregulation of some key genes in the absence of SA2 to evade proper differentiation. Functional studies in the hematopoietic system have indeed shown that reduced cohesin activity promotes transformation by altering the balance between self-renewal and differentiation (Mazumdar et al., 2015; Mullenders et al., 2015; Viny et al., 2015; Fisher et al., 2017). In these studies, the reduced differentiation potential of hematopoietic precursors is similar after decrease of total cohesin -by depletion of a common subunit- or decrease of SA2. We suggest that overall decrease of cohesin may affect more the functions of cohesin-SA2 due to its more dynamic behaviour.

Both cohesin-SA1 and cohesin-SA2 are indispensable for proper embryonic development (Remeseiro et al., 2012a; M. De Koninck and E. Lapi, unpublished results). However, either one appears to be sufficient for cell viability. As expected from the essential nature of the cohesin complex, a synthetic lethality screen in SA2-deficient cells identified SA1 as the strongest candidate (van der Lelij et al., 2017). Consistent with this result, a mutually exclusive mutation pattern has been observed in tumours between the genes encoding the two variants, as well as an inverse correlation between their expression levels (Liu et al., 2018). Importantly, while the loss of both cohesin complexes abrogates cohesion completely and leads to mitotic failure and eventually cell death, SA1 inactivation does not affect proliferation of SA2 proficient cells (van der Lelij et al., 2017). Thus, it has already been suggested that selective inhibition of cohesin-SA1 has big therapeutic potential in treatment of the patients with STAG2-mutated cancers, under the premise that it would not be deleterious for most non-cancerous tissues (Figure D10). Previous work from our group has shown that mice heterozygous for the STAG1 are viable and fertile although display increased incidence of pancreatic cancer (Remeseiro et al., 2012a). It should be noted that these animals have reduced dosage of the STAG1 from their conception and the observed susceptibility to pancreatic cancer could be a consequence of a cumulative effect of aberrant expression of REG genes involved in proliferation, differentiation and protection from inflammation (Cuadrado et al., 2015). Even though selective targeting of cohesin-SA1 presents a formidable challenge due to high homology between two SA paralogs, we think that future studies should explore the effects of acute ablation of cohesin-SA1 in adult mice. Our work in MCF10A cells suggests that the impact of eliminating SA1 is relatively mild in terms of genome organization and gene regulation.

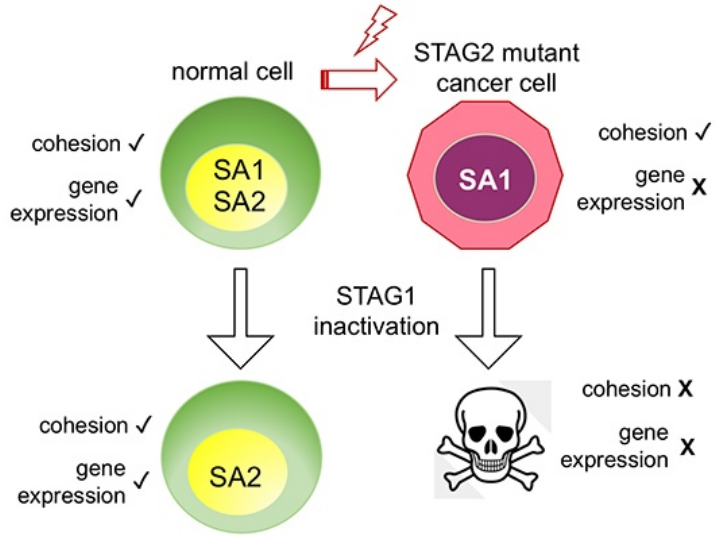


Figure D10. Inactivation of cohesin-SA1 for treatment of STAG2 mutant cancers

See text for details.

A recent study reports the identification of STAG1 microdeletions or point mutations in children affected by intellectual deficiency and variable associated symptoms (Lehalle et al., 2017). Heterozygous mutations in STAG2 and duplications involving STAG2 have also been identified in female and male patients, respectively, with symptoms overlapping those of cohesinopathies (Bonnet et al., 2009; Kumar et al., 2015; Leroy et al., 2016; Philippe et al., 2013; Yingjun et al., 2015). Gene deregulation is likely at the origin of these pathologies, although decreased proliferation or reduced ability to deal with DNA damage in the absence of a full cohesin complement may also contribute. Understanding the role of the two cohesin variants to genome folding and gene regulation in the context of development is a current interest of our research.

Conclusions

1- Cohesin is present throughout the genome at CTCF and non-CTCF positions. Both cohesin-SA1 and cohesin-SA2 are present at CTCF positions (“common”) while cohesin-SA2 is the predominant cohesin variant at non-CTCF cohesin positions (“SA2-only”). Common positions are enriched in insulators whereas SA2-only positions localize mainly in active enhancers.

2- Binding of cohesin to SA2-only positions is more dynamic than binding to common cohesin positions. Cohesin stacking at common positions, observed by Re-ChIP experiments, may hinder the association of cohesin unloading factor WAPL and thereby stabilize cohesin at these binding sites.

3- Cohesin-SA1 cannot be recruited to SA2-only positions in the absence of cohesin-SA2, thus it cannot assume cohesin functions carried out from them. In contrast, the presence of cohesin-SA1 and cohesin-SA2 at common positions depends on CTCF but not on the other variant complex.

4- Cohesin-SA2 contributes to cell-type specific gene expression in a CTCF-independent manner from SA2-only positions through association with different transcriptional regulators. This activity of cohesin-SA2 is in part achieved from cell-type specific super enhancers through regulation of cell-type specific master regulators.

5- Opposite to CTCF, cohesin-SA2 downregulation does not alter overall topology of the S100 locus. However, there is a strong correlation between the presence of SA2-only positions, the intensity of local interactions and the levels of expression of nearby genes.

6- Removal of a single cohesin variant does not alter significantly overall genome organization but affects genomic interactions at different distances. Loss of SA2 increases mid-range (0.1 -1.3 Mb) contacts while loss of SA1 increases long-range (>1.4 Mb) contacts.

7- TAD border strength and TAD border identity depend more on cohesin-SA1 and cohesin-SA2, respectively.

8- Cohesin-SA1 has a more structural role in genome organization, in particular mediating TAD/subTAD formation together with CTCF, while cohesin-SA2 likely facilitates intra-TAD contacts between enhancers and/or promoters, independently of CTCF and together with transcriptional regulators.

Conclusiones

1- El complejo cohesina aparece a lo largo del genoma en posiciones en las que se puede encontrar o no la proteína CTCF. Tanto cohesina-SA1 como cohesina-SA2 se localizan en sitios de unión de CTCF (posiciones "comunes"), mientras que la cohesina-SA2 es la variante más frecuente en los sitios de unión sin CTCF (posiciones "sólo-SA2"). Las primeras se localizan preferentemente en "insulators" y las segundas se encuentran enriquecidas en "enhancers".

2- La unión del complejo cohesina a las posiciones exclusivas de SA2 es más dinámica que su unión a posiciones comunes. La acumulación de más de un complejo de cohesina en una posición determinada, observada en experimentos de Re-ChIP, podría dificultar la unión de WAPL al complejo, evitando su disociación de la cromatina, lo que conduciría a la estabilización de la cohesina en estas posiciones.

3- La cohesina-SA1 no puede ocupar sitios "sólo-SA2" en ausencia de SA2, por lo que no es capaz de desempeñar las funciones asociadas a dichas posiciones. Por el contrario, la presencia de cohesina-SA1 y cohesina-SA2 en las posiciones comunes depende de CTCF y no de la presencia de la otra variante del complejo.

4- La cohesina-SA2 contribuye a la expresión génica específica de cada tipo celular. Esta función depende de su presencia en sitios "sólo-SA2", carentes de CTCF, en los que se asocia con diferentes reguladores de la transcripción.

5- La topología general de la cromatina en el locus S100 depende de CTCF pero no de cohesina-SA2. Existe sin embargo una fuerte correlación entre la presencia de posiciones exclusivas de SA2, la intensidad de interacciones locales y los niveles de expresión de los genes adyacentes.

6- La eliminación de una única variante del complejo cohesina no altera la estructura general del genoma de modo significativo. Sin embargo, provoca cambios en las interacciones genómicas a diferentes distancias: la pérdida de SA2 aumenta los contactos a distancias medias (de 0.1 a 1.3 megabases), mientras que la pérdida de SA1 aumenta las interacciones a largas distancias (>1.4 Mb).

7- La fortaleza y la identidad de los bordes de los TADs dependen en mayor medida de cohesina-SA1 y cohesina-SA2, respectivamente.

8- Mientras que la cohesina-SA1 desempeña mayoritariamente un papel estructural en la organización del genoma, mediando la formación de TADs y subTADs junto con CTCF, la

cohesina-SA2 facilita la formación de contactos intra-TAD entre “enhancers” y promotores, de modo independiente de CTCF y en estrecha relación con distintos reguladores de la transcripción.

Bibliography

Anderson, D.E., Losada, A., Erickson, H.P., and Hirano, T. (2002). Condensin and cohesin display different arm conformations with characteristic hinge angles. J. Cell Biol. *156*, 419–424.

Antony, J., Dasgupta, T., Rhodes, J.M., McEwan, M. V., Print, C.G., O'Sullivan, J.M., and Horsfield, J.A. (2015). Cohesin modulates transcription of estrogen-responsive genes. Biochim. Biophys. Acta - Gene Regul. Mech. *1849*, 257–269.

Arumugam, P., Gruber, S., Tanaka, K., Haering, C.H., Mechtler, K., and Nasmyth, K. (2003). ATP Hydrolysis Is Required for Cohesin's Association with Chromosomes. Curr. Biol. *13*, 1941–1953.

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME Suite: Tools for motif discovery and searching. Nucleic Acids Res. *37*.

Balbás-Martínez, C., Sagrera, A., Carrillo-De-Santa-Pau, E., Earl, J., Márquez, M., Vazquez, M., Lapi, E., Castro-Giner, F., Beltran, S., Bayés, M., et al. (2013). Recurrent inactivation of STAG2 in bladder cancer is not associated with aneuploidy. Nat. Genet. *45*, 1464–1469.

Ballas, N., Grunseich, C., Lu, D.D., Speh, J.C., and Mandel, G. (2005). REST and its corepressors mediate plasticity of neuronal gene chromatin throughout neurogenesis. Cell *121*, 645–657.

Banerji, R., Skibbens, R. V., and Iovine, M.K. (2017a). How many roads lead to cohesinopathies? Dev. Dyn. *246*, 881–888.

Barber, T.D., McManus, K., Yuen, K.W., Reis, M., Parmigiani, G., Shen, D., Barrett, I., Nouhi, Y., Spencer, F., Markowitz, S., et al. (2008). Chromatid cohesion defects may underlie chromosome instability in human colorectal cancers. Proc Natl Acad Sci U S A *105*, 3443–3448.

Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-Resolution Profiling of Histone Methylations in the Human Genome. Cell *129*, 823–837.

Bell, A.C., and Felsenfeld, G. (2000). Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. Nature *405*, 482–485.

Bianco, S., Chiariello, A.M., Annunziatella, C., Esposito, A., and Nicodemi, M. (2017). Predicting chromatin architecture from models of polymer physics. Chromosom. Res. *25*, 25–34.

Bisht, K.K., Daniloski, Z., and Smith, S. (2013). SA1 binds directly to DNA through its unique AT-hook to promote sister chromatid cohesion at telomeres. J. Cell Sci. *126*, 3493–3503.

Boettiger, A.N., Bintu, B., Moffitt, J.R., Wang, S., Beliveau, B.J., Fudenberg, G., Imakaev, M., Mirny, L.A., Wu, C.T., and Zhuang, X. (2016). Super-resolution imaging reveals distinct chromatin folding for different epigenetic states. Nature *529*, 418–422.

Bonev, B., and Cavalli, G. (2016). Organization and function of the 3D genome. *Nature Reviews Genetics.* 17, 661–678.

Bonev, B., Cohen, N.M., Szabo, Q., Hugnot, J.-P., Tanay, A., Cavalli, G., Bonev, B., Cohen, N.M., Szabo, Q., Fritsch, L., et al. (2017). Multiscale 3D Genome Rewiring during Mouse Article Multiscale 3D Genome Rewiring during Mouse Neural Development. Cell *171*, 557.e1-557.e24.

Bonnet, C., Leheup, B., Béri, M., Philippe, C., Grégoire, M.J., and Jonveaux, P. (2009). Aberrant GRIA3 transcripts with multi-exon duplications in a family with X-linked mental retardation. Am. J. Med. Genet. Part A *149*, 1280–1289.

Brohl, A.S., Solomon, D.A., Chang, W., Wang, J., Song, Y., Sindiri, S., Patidar, R., Hurd, L., Chen, L., Shern, J.F., et al. (2014). The Genomic Landscape of the Ewing Sarcoma Family of Tumors Reveals Recurrent STAG2 Mutation. PLoS Genet. *10*.

Buheitel, J., and Stemmann, O. (2013). Prophase pathway-dependent removal of cohesin from human chromosomes requires opening of the Smc3-Scc1 gate. EMBO J. *32*, 666–676.

Busslinger, G.A., Stocsits, R.R., Van Der Lelij, P., Axelsson, E., Tedeschi, A., Galjart, N., and Peters, J.M. (2017). Cohesin is positioned in mammalian genomes by transcription, CTCF and WAPL. Nature *544*, 503–507.

Canudas, S., Houghtaling, B.R., Kim, J.Y., Dynek, J.N., Chang, W.G., and Smith, S. (2007). Protein requirements for sister telomere association in human cells. EMBO J. *26*, 4867–4878.

Canudas, S., and Smith, S. (2009). Differential regulation of telomere and centromere cohesion by the Scc3 homologues SA1 and SA2, respectively, in human cells. J. Cell Biol. *187*, 165–173.

Chao, W.C.H., Murayama, Y., Muñoz, S., Costa, A., Uhlmann, F., and Singleton, M.R. (2015). Structural Studies Reveal the Functional Modularity of the Scc2-Scc4 Cohesin Loader. Cell Rep. *12*, 719–725.

Chao, W.C.H., Murayama, Y., Muñoz, S., Jones, A.W., Wade, B.O., Purkiss, A.G., Hu, X.W., Borg, A., Snijders, A.P., Uhlmann, F., et al. (2017). Structure of the cohesin loader Scc2. Nat. Commun. *8*.

Countryman, P., Fan, Y., Gorthi, A., Pan, H., Strickland, J., Kaur, P., Wang, X., Lin, J., Lei, X., White, C., et al. (2018). Cohesin SA2 is a sequence-independent DNA-binding protein that recognizes DNA replication and repair intermediates. J. Biol. Chem. *293*, 1054–1069.

Cremer, T., and Cremer, C. (2001). Chromosome territories, nuclear architecture and gene regulation in mammalian cells. Nat. Rev. Genet. *2*, 292–301.

Crompton, B.D., Stewart, C., Taylor-Weiner, A., Alexe, G., Kurek, K.C., Calicchio, M.L., Kiezun, A., Carter, S.L., Shukla, S.A., Mehta, S.S., et al. (2014). The genomic landscape of pediatric Ewing sarcoma. Cancer Discov. *4*, 1326–1341.

D'Alessio, A.C., Fan, Z.P., Wert, K.J., Baranov, P., Cohen, M.A., Saini, J.S., Cohick, E., Charniga, C., Dadon, D., Hannett, N.M., et al. (2015). A systematic approach to identify candidate transcription factors that control cell identity. Stem Cell Reports *5*, 763–775.

Davidson, I.F., Goetz, D., Zaczek, M.P., Molodtsov, M.I., Huis in 't Veld, P.J., Weissmann, F., Litos, G., Cisneros, D.A., Ocampo-Hafalla, M., Ladurner, R., et al. (2016). Rapid movement and transcriptional re-localization of human cohesin on DNA. EMBO J. *35*, 2671–2685.

De Koninck, M., and Losada, A. (2016). Cohesin mutations in cancer. Cold Spring Harb. Perspect. Med. *6*.

De Laat, W., and Duboule, D. (2013). Topology of mammalian developmental enhancers and their regulatory landscapes. Nature *502*, 499–506.

de Wit, E., and de Laat, W. (2012). A decade of 3C technologies-insights into nuclear organization. Genes Dev. 11–24.

de Wit, E., Vos, E.S.M., Holwerda, S.J.B., Valdes-Quezada, C., Verstegen, M.J.A.M., Teunissen, H., Splinter, E., Wijchers, P.J., Krijger, P.H.L., and de Laat, W. (2015). CTCF Binding Polarity Determines Chromatin Looping. Mol. Cell *60*, 676–684.

Deardorff, M.A., Bando, M., Nakato, R., Watrin, E., Itoh, T., Minamino, M., Saitoh, K., Komata, M., Katou, Y., Clark, D., et al. (2012). HDAC8 mutations in Cornelia de Lange syndrome affect the cohesin acetylation cycle. Nature *489*, 313–317.

Degner, S.C., Verma-Gaur, J., Wong, T.P., Bossen, C., Iverson, G.M., Torkamani, A., Vettermann, C., Lin, Y.C., Ju, Z., Schulz, D., et al. (2011). CCCTC-binding factor (CTCF) and cohesin influence the genomic architecture of the Igh locus and antisense transcription in pro-B cells. Proc. Natl. Acad. Sci. *108*, 9566–9571.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. Science (80-.). *295*, 1306–1311.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380.

Dixon, J.R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J.E., Lee, A.Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W., et al. (2015). Chromatin architecture reorganization during stem cell differentiation. Nature *518*, 331–336.

Dolgin, E. (2017). DNA's secret weapon against knots and tangles. Nature *544*, 284–286.

Ea, V., Baudement, M.O., Lesne, A., and Forné, T. (2015). Contribution of topological domains and loop formation to 3D chromatin organization. Genes (Basel). *6*, 734–750.

Ernst, J. et al, Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43–49.

Fabre, P.J., Benke, A., Joye, E., Nguyen Huynh, T.H., Manley, S., and Duboule, D. (2015). Nanoscale spatial organization of the *HoxD* gene cluster in distinct transcriptional states. Proc. Natl. Acad. Sci. *112*, 13964–13969.

Faure, A.J., Schmidt, D., Watt, S., Schwalie, P.C., Wilson, M.D., Xu, H., Ramsay, R.G., Odom, D.T., and Flicek, P. (2012). Cohesin regulates tissue-specific expression by stabilizing highly occupied cis-regulatory modules. Genome Res. *22*, 2163–2175.

Fisher, J.B., Peterson, J., Reimer, M., Stelloh, C., Pulakanti, K., Gerbec, Z.J., Abel, A.M., Strouse, J.M., Strouse, C., McNulty, M., et al. (2017). The cohesin subunit Rad21 is a negative regulator of hematopoietic self-renewal through epigenetic repression of Hoxa7 and Hoxa9. Leukemia *31*, 712–719.

Flavahan, W.A., Drier, Y., Liau, B.B., Gillespie, S.M., Venteicher, A.S., Stemmer-Rachamimov, A.O., Suvà, M.L., and Bernstein, B.E. (2016). Insulator dysfunction and oncogene activation in IDH mutant gliomas. Nature *529*, 110–114.

Flyamer, I.M., Gassler, J., Imakaev, M., Brandão, H.B., Ulianov, S. V., Abdennur, N., Razin, S. V., Mirny, L.A., and Tachibana-Konwalski, K. (2017). Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. Nature *544*, 110–114.

Franke, M., Ibrahim, D.M., Andrey, G., Schwarzer, W., Heinrich, V., Schï¿½pflin, R., Kraft, K., Kempfer, R., Jerković, I., Chan, W.L., et al. (2016). Formation of new chromatin domains determines pathogenicity of genomic duplications. Nature *538*, 265–269.

Fraser, J., Ferrai, C., Chiariello, A.M., Schueler, M., Rito, T., Laudanno, G., Barbieri, M., Moore, B.L., Kraemer, D.C., Aitken, S., et al. (2015). Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. Mol. Syst. Biol. *11*, 852–852.

Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N., and Mirny, L.A. (2016). Formation of Chromosomal Domains by Loop Extrusion. Cell Rep. *15*, 2038–2049.

Fussner, E., Strauss, M., Djuric, U., Li, R., Ahmed, K., Hart, M., Ellis, J., and Bazett-Jones, D.P. (2012). Open and closed domains in the mouse genome are configured as 10-nm chromatin fibres. EMBO Rep. *13*, 992–996.

Gandhi, R., Gillespie, P.J., and Hirano, T. (2006). Human WAPL Is a Cohesin-Binding Protein that Promotes Sister-Chromatid Resolution in Mitotic Prophase. Curr. Biol. *16*, 2406–2417.

Ganji, M., Shaltiel, I.A., Bisht, S., Kim, E., Kalichava, A., Haering, C.H., and Dekker, C. (2018). Real-time imaging of DNA loop extrusion by condensin. Science. *360*, 102–105.

Gassler, J., Brandão, H.B., Imakaev, M., Flyamer, I.M., Ladstätter, S., Bickmore, W.A., Peters, J., Mirny, L.A., and Tachibana, K. (2017). A mechanism of cohesin-dependent loop extrusion organizes zygotic genome architecture. EMBO J. e201798083.

Gerlich, D., Koch, B., Dupeux, F., Peters, J.M., and Ellenberg, J. (2006). Live-cell imaging reveals a stable cohesin-chromatin interaction after but not before DNA replication . Curr. Biol. *16*, 1571–1578.

Gibcus, J.H., and Dekker, J. (2013). The Hierarchy of the 3D Genome. Mol. Cell *49*, 773–782.

Gligoris, T.G., Scheinost, J.C., B??rmann, F., Petela, N., Chan, K.L., Uluocak, P., Beckouet, F., Gruber, S., Nasmyth, K., and Lowe, J. (2014). Closing the cohesin ring: Structure and function of its Smc3-kleisin interface. Science. *346*, 963–967.

Glynn, E.F., Megee, P.C., Yu, H.G., Mistrot, C., Unal, E., Koshland, D.E., DeRisi, J.L., and Gerton, J.L. (2004). Genome-wide mapping of the cohesin complex in the yeast Saccharomyces cerevisiae. PLoS Biol. *2*.

Gruber, S., Haering, C.H., and Nasmyth, K. (2003). Chromosomal cohesin forms a ring. Cell *112*, 765–777.

Gruber, S., Arumugam, P., Katou, Y., Kuglitsch, D., Helmhart, W., Shirahige, K., and Nasmyth, K. (2006). Evidence that Loading of Cohesin Onto Chromosomes Involves Opening of Its SMC Hinge. Cell *127*, 523–537.

Guacci, V., Koshland, D., and Strunnikov, A. (1997). A direct link between sister chromatid cohesion and chromosome condensation revealed through the analysis of MCD1 in S. cerevisiae. Cell *91*, 47–57.

Guillou, E., Ibarra, A., Coulon, V., Casado-Vela, J., Rico, D., Casal, I., Schwob, E., Losada, A., and Méndez, J. (2010). Cohesin organizes chromatin loops at DNA replication factories. Genes Dev. *24*, 2812–2822.

Guo, Y., Monahan, K., Wu, H., Gertz, J., Varley, K.E., Li, W., Myers, R.M., Maniatis, T., and Wu, Q. (2012). CTCF/cohesin-mediated DNA looping is required for protocadherin promoter choice. Proc. Natl. Acad. Sci. *109*, 21081–21086.

Guo, G., Sun, X., Chen, C., Wu, S., Huang, P., Li, Z., Dean, M., Huang, Y., Jia, W., Zhou, Q., et al. (2013). Whole-genome and whole-exome sequencing of bladder cancer identifies frequent alterations in genes involved in sister chromatid cohesion and segregation. Nat. Genet. *45*, 1459–1463.

Guo, Y., Xu, Q., Canzio, D., Shou, J., Li, J., Gorkin, D.U., Jung, I., Wu, H., Zhai, Y., Tang, Y., et al. (2015). CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. Cell *162*, 900–910.

Haarhuis, J.H.I., van der Weide, R.H., Blomen, V.A., Yáñez-Cuna, J.O., Amendola, M., van Ruiten, M.S., Krijger, P.H.L., Teunissen, H., Medema, R.H., van Steensel, B., et al. (2017). The Cohesin Release Factor WAPL Restricts Chromatin Loop Extension. Cell *169*, 693–707.e14.

Haddad, N., Jost, D., and Vaillant, C. (2017). Perspectives: using polymer modeling to understand the formation and function of nuclear compartments. Chromosom. Res. *25*, 35–50.

Hadjur, S., Williams, L.M., Ryan, N.K., Cobb, B.S., Sexton, T., Fraser, P., Fisher, A.G., and Merkenschlager, M. (2009). Cohesins form chromosomal cis-interactions at the developmentally regulated IFNG locus. Nature *460*, 410–413.

Haering, C.H., Farcas, A.M., Arumugam, P., Metson, J., and Nasmyth, K. (2008). The cohesin ring concatenates sister DNA molecules. Nature *454*, 297–301.

Haering, C.H., Löwe, J., Hochwagen, A., and Nasmyth, K. (2002). Molecular architecture of SMC proteins and the yeast cohesin complex. Mol. Cell *9*, 773–788.

Hara, K., Zheng, G., Qu, Q., Liu, H., Ouyang, Z., Chen, Z., Tomchick, D.R., and Yu, H. (2014). Structure of cohesin subcomplex pinpoints direct shugoshin-WAPL antagonism in centromeric cohesion. Nat. Struct. Mol. Biol. *21*, 864–870.

Hansen, A.S., Cattoglio, C., Darzacq, X., and Tjian, R. (2017). Recent evidence that TADs and chromatin loops are dynamic structures. Nucleus *1034*, 1–13.

Hark, A.T., Schoenherr, C.J., Katz, D.J., Ingram, R.S., Levorse, J.M., and Tilghman, S.M. (2000). CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. Nature *405*, 486–489.

Hashimoto, H., Wang, D., Horton, J.R., Zhang, X., Corces, V.G., and Cheng, X. (2017). Structural Basis for the Versatile and Methylation-Dependent Binding of CTCF to DNA. Mol. Cell *66*, 711–720.e3.

Heidinger-Pauli, J.M., Mert, O., Davenport, C., Guacci, V., and Koshland, D. (2010). Systematic Reduction of Cohesin Differentially Affects Chromosome Segregation, Condensation, and DNA Repair. Curr. Biol. *20*, 957–963.

Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-André, V., Sigova, A.A., Hoke, H.A., and Young, R.A. (2013). Super-enhancers in the control of cell identity and disease. Cell *155*, 934–947.

Hnisz, D., Day, D.S., and Young, R.A. (2016). Insulated Neighborhoods: Structural and Functional Units of Mammalian Gene Control. Cell *167*, 1188–1200.

Horsfield, J.A., Anagnostou, S.H., Hu, J.K.-H., Cho, K.H.Y., Geisler, R., Lieschke, G., Crosier, K.E., and Crosier, P.S. (2007). Cohesin-dependent regulation of Runx genes. Development *134*, 2639–2649.

Hu, B., Itoh, T., Mishra, A., Katoh, Y., Chan, K.L., Upcher, W., Godlee, C., Roig, M.B., Shirahige, K., and Nasmyth, K. (2011). ATP hydrolysis is required for relocating cohesin from sites occupied by its Scc2/4 loading complex. Curr. Biol. *21*, 12–24.

Hughes, J.R., Roberts, N., Mcgowan, S., Hay, D., Giannoulatou, E., Lynch, M., De Gobbi, M., Taylor, S., Gibbons, R., and Higgs, D.R. (2014). Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. Nat. Genet. *46*, 205–212.

Huis In't Veld, P.J., Herzog, F., Ladurner, R., Davidson, I.F., Piric, S., Kreidl, E., Bhaskara, V., Aebersold, R., and Peters, J.M. (2014). Characterization of a DNA exit gate in the human cohesin ring. Science (80-.). *346*, 968–972.

Izumi, K., Nakato, R., Zhang, Z., Edmondson, A., Noon, S., Dulik, M., Rajagopalan, R., Venditti, C., Gripp, K., Samanich, J., Zackai, E., Deardorff, M., Clark, D., Allen, J., Dorsett, D., Misulovin, Z., Komata, M., Bando, M., Kaur, M., Katou, Y., Shirahige, K. and Krantz, I. (2015). Germline gain-of-function mutations in AFF4 cause a developmental syndrome functionally linking the super elongation complex and cohesin. *Nature Genetics*, *47*:338-344.

Jackson, D.A., and Pombo, A. (1998). Replicon clusters are stable units of chromosome structure: Evidence that nuclear organization contributes to the efficient activation and propagation of S phase in human cells. J. Cell Biol. *140*, 1285–1295.

Jan, M., Snyder, T.M., Corces-Zimmerman, M.R., Vyas, P., Weissman, I.L., Quake, S.R., and Majeti, R. (2012). Clonal evolution of preleukemic hematopoietic stem cells precedes human acute myeloid leukemia. Sci. Transl. Med. *4*.

Jang, W., Kim, T., Koo, J.S., Kim, S., and Lim, D. (2017). Mechanical cue-induced YAP instructs Skp2-dependent cell cycle exit and oncogenic signaling. EMBO J. e201696089.

Jeppsson, K., Carlborg, K.K., Nakato, R., Berta, D.G., Lilienthal, I., Kanno, T., Lindqvist, A., Brink, M.C., Dantuma, N.P., Katou, Y., et al. (2014). The Chromosomal Association of the Smc5/6 Complex Depends on Cohesion and Predicts the Level of Sister Chromatid Entanglement. PLoS Genet. *10*.

Jin, F., Li, Y., Dixon, J.R., Selvaraj, S., Ye, Z., Lee, A.Y., Yen, C.A., Schmitt, A.D., Espinoza, C.A., and Ren, B. (2013). A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature *503*, 290–294.

Kagey, M.H., Newman, J.J., Bilodeau, S., Zhan, Y., Orlando, D.A., Van Berkum, N.L., Ebmeier, C.C., Goossens, J., Rahl, P.B., Levine, S.S., et al. (2010). Mediator and cohesin connect gene expression and chromatin architecture. Nature *467*, 430–435.

Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J.F., Wyczalkowski, M.A., et al. (2013). Mutational landscape and significance across 12 major cancer types. Nature *502*, 333–339.

Kanke, M., Tahara, E., Huis in't Veld, P.J., and Nishiyama, T. (2016). Cohesin acetylation and WAPL-Pds5 oppositely regulate translocation of cohesin along DNA. EMBO J. *35*, 2686–2698.

Katainen, R., Dave, K., Pitkänen, E., Palin, K., Kivioja, T., Välimäki, N., Gylfe, A.E., Ristolainen, H., Hänninen, U.A., Cajuso, T., et al. (2015). CTCF/cohesin-binding sites are frequently mutated in cancer. Nat. Genet. *47*, 818–821.

Kawauchi, S., Calof, A.L., Santos, R., Lopez-Burks, M.E., Young, C.M., Hoang, M.P., Chua, A., Lao, T., Lechner, M.S., Daniel, J.A., et al. (2009). Multiple organ system defects and transcriptional dysregulation in the Nipbl+/- mouse, a model of Cornelia de Lange syndrome. PLoS Genet. *5*.

Kikuchi, S., Borek, D.M., Otwinowski, Z., Tomchick, D.R., and Yu, H. (2016). Crystal structure of the cohesin loader Scc2 and insight into cohesinopathy. Proc. Natl. Acad. Sci. *113*, 12444–12449.

Kim, T.H., Abdullaev, Z.K., Smith, A.D., Ching, K.A., Loukinov, D.I., Green, R.D.D., Zhang, M.Q., Lobanenkov, V. V., and Ren, B. (2007). Analysis of the Vertebrate Insulator Protein CTCF-Binding Sites in the Human Genome. Cell *128*, 1231–1245.

Kim, J.S., He, X., Orr, B., Wutz, G., Hill, V., Peters, J.M., Compton, D.A., and Waldman, T. (2016). Intact Cohesion, Anaphase, and Chromosome Segregation in Human Cells Harboring Tumor-Derived Mutations in STAG2. PLoS Genet. *12*.

Kon, A., Shih, L.Y., Minamino, M., Sanada, M., Shiraishi, Y., Nagata, Y., Yoshida, K., Okuno, Y., Bando, M., Nakato, R., et al. (2013). Recurrent mutations in multiple components of the cohesin complex in myeloid neoplasms. Nat. Genet. *45*, 1232–1237.

Kouzarides, T. (2007). Chromatin Modifications and Their Function. Cell *128*, 693–705.

Kumar, R., Corbett, M.A., Van Bon, B.W.M., Gardner, A., A.Woenig, J., Jolly, L.A., Douglas, E., Friend, K., Tan, C., Van Esch, H., et al. (2015). Increased STAG2 dosage defines a novel cohesinopathy with intellectual disability and behavioral problems. Hum. Mol. Genet. *24*, 7171–7181.

Kurukuti, S., Tiwari, V.K., Tavoosidana, G., Pugacheva, E., Murrell, A., Zhao, Z., Lobanenkov, V., Reik, W., and Ohlsson, R. (2006). CTCF binding at the H19 imprinting

control region mediates maternally inherited higher-order chromatin conformation to restrict enhancer access to Igf2. Proc. Natl. Acad. Sci. *103*, 10684–10689.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature *499*, 214–218.

Le Dily, F.L., Baù, D., Pohl, A., Vicent, G.P., Serra, F., Soronellas, D., Castellano, G., Wright, R.H.G., Ballare, C., Filion, G., et al. (2014). Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation. Genes Dev. *28*, 2151–2162.

Leiserson, M.D.M., Vandin, F., Wu, H.T., Dobson, J.R., Eldridge, J. V., Thomas, J.L., Papoutsaki, A., Kim, Y., Niu, B., McLellan, M., et al. (2015). Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nat. Genet. *47*, 106–114.

Lehalle, D., Mosca-Boidron, A.L., Begtrup, A., Boute-Benejean, O., Charles, P., Cho, M.T., Clarkson, A., Devinsky, O., Duffourd, Y., Duplomb-Jego, L., et al. (2017). STAG1 mutations cause a novel cohesinopathy characterised by unspecific syndromic intellectual disability. J. Med. Genet. *54*, 479–488.

Lemaitre, J.M., Danis, E., Pasero, P., Vassetzky, Y., and Méchali, M. (2005). Mitotic remodeling of the replicon and chromosome structure. Cell *123*, 787–801.

Lengronne, A., Katou, Y., Mori, S., Yokabayashi, S., Kelly, G.P., Ito, T., Watanabe, Y., Shirahige, K., and Uhlmann, F. (2004). Cohesin relocation from sites of chromosomal loading to places of convergent transcription. Nature *430*, 573–578.

Leroy, C., Jacquemont, M.L., Doray, B., Lamblin, D., Cormier-Daire, V., Philippe, A., Nusbaum, S., Patrat, C., Steffann, J., Colleaux, L., et al. (2016). Xq25 duplication: The crucial role of the STAG2 gene in this novel human cohesinopathy. Clin. Genet. *89*, 68–73.

Li, Z., Zhang, P., Yan, A., Guo, Z., Ban, Y., Li, J., Chen, S., Yang, H., He, Y., Li, J., et al. (2017). ASXL1 interacts with the cohesin complex to maintain chromatid separation and gene expression for normal hematopoiesis. Sci. Adv. *3*.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

Lieberman-Aiden, E., Van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. *326*, 289–293.

Lin, Z., Luo, X., and Yu, H. (2016). Structural basis of cohesin cleavage by separase. Nature *532*, 131–134.

Lin, J., Countryman, P., Chen, H., Pan, H., Fan, Y., Jiang, Y., Kaur, P., Miao, W., Gurgel, G., You, C., et al. (2016a). Functional interplay between SA1 and TRF1 in telomeric DNA binding and DNA-DNA pairing. Nucleic Acids Res. *44*, 6363–6376.

Liu, H., Rankin, S., and Yu, H. (2013). Phosphorylation-enabled binding of SGO1-PP2A to cohesin protects sororin and centromeric cohesion during mitosis. Nat. Cell Biol. *15*, 40–49.

Liu, J., Feldman, R., Zhang, Z., Deardorff, M.A., Haverfield, E. V., Kaur, M., Li, J.R., Clark, D., Kline, A.D., Waggoner, D.J., et al. (2009). SMC1A expression and mechanism of pathogenicity in probands with X-linked Cornelia de Lange Syndrome. Hum. Mutat. *30*, 1535–1542.

Liu, Y., Xu, H., Van der Jeught, K., Li, Y., Liu, S., Zhang, L., Fang, Y., Radovich, M., Schneider, B.P., He, X., Huang, C., et al. (2018). Somatic mutation of the cohesin complex subunit confers therapeutic vulnerabilities in cancer. J Clin Invest. https://doi.org/10.1172/JCI98727.

Lobanenkov, V. V, Nicolas, R.H., Adler, V. V, Paterson, H., Klenova, E.M., Polotskaja, A. V, and Goodwin, G.H. (1990). A novel sequence-specific DNA binding protein which interacts with three regularly spaced direct repeats of the CCCTC-motif in the 5'-flanking sequence of the chicken c-myc gene. Oncogene *5*, 1743–1753.

Lonfat, N., Montavon, T., Darbellay, F., Gitto, S., and Duboule, D. (2014). Convergent evolution of complex regulatory landscapes and pleiotropy at Hox loci. Science. *346*, 1004–1006.

Losada, A., Hirano, M., and Hirano, T. (1998). Identification of Xenopus SMC protein complexes required for sister chromatid cohesion. Genes Dev. *12*, 1986–1997.

Losada, A., and Hirano, T. (2001). Intermolecular DNA interactions stimulated by the cohesin complex in vitro: Implications for sister chromatid cohesion. Curr. Biol. *11*, 268–272.

Losada, A., Hirano, M., and Hirano, T. (2002). Cohesin release is required for sister chromatid resolution, but not for condensin-mediated compaction, at the onset of mitosis. Genes Dev. *16*, 3004–3016.

Losada, A., Yokochi, T., Kobayashi, R., and Hirano, T. (2000). Identification and characterization of SA/Scc3p subunits in the Xenopus and human cohesin complexes. J. Cell Biol. *150*, 405–416.

Lupiáñez, D.G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J.M., Laxova, R., et al. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell *161*, 1012–1025.

Machanick, P., and Bailey, T.L. (2011). MEME-ChIP: Motif analysis of large DNA datasets. Bioinformatics *27*, 1696–1697.

Mammana, A., Chung, H.R. (2015). Chromatin segmentation based on a probabilistic model for read counts explains a large portion of the genome. Genome Biol. 16, 151.

Martincorena, I., Raine, K.M., Gerstung, M., Dawson, K.J., Haase, K., Van Loo, P., Davies, H., Stratton, M.R., and Campbell, P.J. (2017). Universal Patterns of Selection in Cancer and Somatic Tissues. Cell *171*, 1029–1041.

Mazumdar, C., Shen, Y., Xavy, S., Zhao, F., Reinisch, A., Li, R., Corces, M.R., Flynn, R.A., Buenrostro, J.D., Chan, S.M., et al. (2015). Leukemia-Associated Cohesin Mutants Dominantly Enforce Stem Cell Programs and Impair Human Hematopoietic Progenitor Differentiation. Cell Stem Cell *17*, 675–688.

Michaelis, C., Ciosk, R., and Nasmyth, K. (1997). Cohesins: Chromosomal proteins that prevent premature separation of sister chromatids. Cell *91*, 35–45.

Mifsud, B., Tavares-Cadete, F., Young, A.N., Sugar, R., Schoenfelder, S., Ferreira, L., Wingett, S.W., Andrews, S., Grey, W., Ewels, P.A., et al. (2015). Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nat. Genet. *47*, 598–606.

Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.K., Koche, R.P., et al. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature *448*, 553–560.

Mishiro, T., Ishihara, K., Hino, S., Tsutsumi, S., Aburatani, H., Shirahige, K., Kinoshita, Y., and Nakao, M. (2009). Architectural roles of multiple chromatin insulators at the human apolipoprotein gene cluster. EMBO J. *28*, 1234–1245.

Misulovin, Z., Schwartz, Y.B., Li, X.Y., Kahn, T.G., Gause, M., MacArthur, S., Fay, J.C., Eisen, M.B., Pirrotta, V., Biggin, M.D., et al. (2008). Association of cohesin and Nipped-B with transcriptionally active regions of the Drosophila melanogaster genome. Chromosoma *117*, 89–102.

Monahan, K., Rudnick, N.D., Kehayova, P.D., Pauli, F., Newberry, K.M., Myers, R.M., and Maniatis, T. (2012). Role of CCCTC binding factor (CTCF) and cohesin in the generation of single-cell diversity of Protocadherin-beta gene expression. Proc. Natl. Acad. Sci. *109*, 9125–9130.

Morales, C., and Losada, A. (2018). Establishing and dissolving cohesion during the vertebrate cell cycle. Curr. Opin. Cell Biol. *52*, 51–57.

Mullenders, J., Aranda-Orgilles, B., Lhoumaud, P., Keller, M., Pae, J., Wang, K., Kayembe, C., Rocha, P.P., Raviram, R., Gong, Y., et al. (2015). Cohesin loss alters adult hematopoietic stem cell homeostasis, leading to myeloproliferative neoplasms. J. Exp. Med. *212*, 1833–1850.

Murayama, Y., Samora, C.P., Kurokawa, Y., Iwasaki, H., and Uhlmann, F. (2018). Establishment of DNA-DNA Interactions by the Cohesin Ring. Cell *172*, 465–477.e15.

Murayama, Y., and Uhlmann, F. (2014). Biochemical reconstitution of topological DNA binding by the cohesin ring. Nature *505*, 367–371.

Nasmyth, K. (2001). Disseminating the Genome: Joining, Resolving, and Separating Sister Chromatids During Mitosis and Meiosis. Annu. Rev. Genet. *35*, 673–745.

Nasmyth, K., and Haering, C.H. (2009). Cohesin: Its Roles and Mechanisms. Annu. Rev. Genet. *43*, 525–558.

Nasmyth, K., Peters, J.M., and Uhlmann, F. (2000). Splitting the chromosome: Cutting the ties that bind sister chromatids. Science *288,* 1379-1384.

Natsume, T., Kiyomitsu, T., Saga, Y., and Kanemaki, M.T. (2016). Rapid Protein Depletion in Human Cells by Auxin-Inducible Degron Tagging with Short Homology Donors. Cell Rep. *15*, 210–218.

Nishimura, K., Fukagawa, T., Takisawa, H., Kakimoto, T., and Kanemaki, M. (2009). An auxin-based degron system for the rapid depletion of proteins in nonplant cells. Nat. Methods *6*, 917–922.

Nishiyama, T., Ladurner, R., Schmitz, J., Kreidl, E., Schleiffer, A., Bhaskara, V., Bando, M., Shirahige, K., Hyman, A.A., Mechtler, K., et al. (2010). Sororin mediates sister chromatid cohesion by antagonizing WAPL. Cell *143*, 737–749.

Nishiyama, T., Sykora, M.M., Huis in ’t Veld, P.J., Mechtler, K., and Peters, J.-M. (2013). Aurora B and Cdk1 mediate WAPL activation and release of acetylated cohesin from chromosomes by phosphorylating Sororin. Proc. Natl. Acad. Sci. *110*, 13404–13409.

Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., Van Berkum, N.L., Meisig, J., Sedat, J., et al. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature *485*, 381–385.

Nora, E.P., Goloborodko, A., Valton, A.L., Gibcus, J.H., Uebersohn, A., Abdennur, N., Dekker, J., Mirny, L.A., and Bruneau, B.G. (2017). Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. Cell. *169*, 930–944.e22.

Norton, H.K., and Phillips-Cremins, J.E. (2017). Crossed wires: 3D genome misfolding in human disease. J. Cell Biol. *216*, 3441–3452.

Ocampo-Hafalla, M.T., and Uhlmann, F. (2011). Cohesin loading and sliding. J. Cell Sci. *124*, 685–691.

Ong, C.T., and Corces, V.G. (2014). CTCF: An architectural protein bridging genome topology and function. Nat. Rev. Genet. *15*, 234–246.

Ouyang, Z., Zheng, G., Tomchick, D.R., Luo, X., and Yu, H. (2016). Structural Basis and IP6 Requirement for Pds5-Dependent Cohesin Dynamics. Mol. Cell. *62*, 248–259.

Parelho, V., Hadjur, S., Spivakov, M., Leleu, M., Sauer, S., Gregson, H.C., Jarmuz, A., Canzonetta, C., Webster, Z., Nesterova, T., et al. (2008). Cohesins Functionally Associate with CTCF on Mammalian Chromosome Arms. Cell *132*, 422–433.

Petela, N., Gligoris, T.G., Metson, J.S., Lee, B.-G., Voulgaris, M., Hu, B., Kikuchi, S., Chapard, C., Chen, W., Rajendra, E., et al. (2017). Multiple interactions between Scc1 and Scc2 activate cohesin's DNA dependent ATPase and replace Pds5 during loading. BioRxiv 205914.

Pauli, A., Althoff, F., Oliveira, R.A., Heidmann, S., Schuldiner, O., Lehner, C.F., Dickson, B.J., and Nasmyth, K. (2008). Cell-Type-Specific TEV Protease Cleavage Reveals Cohesin Functions in Drosophila Neurons. Dev. Cell *14*, 239–251.

Pezic, D., Weeks, S.L., Hadjur, S. (2017). More to cohesin than meets the eye: complex diversity for fine-tuning of function. Curr Opin Genet Dev. 43, 93-100.

Philippe, A., Malan, V., Jacquemont, M.L., Boddaert, N., Bonnefont, J.P., Odent, S., Munnich, A., Colleaux, L., and Cormier-Daire, V. (2013). Xq25 duplications encompassing GRIA3 and STAG2 genes in two families convey recognizable X-linked intellectual disability with distinctive facial appearance. Am. J. Med. Genet. Part A *161*, 1370–1375.

Pueschel, R., Coraggio, F., and Meister, P. (2016). From single genes to entire genomes: the search for a function of nuclear organization. Development *143*, 910–923.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. *44*, W160–W165.

Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., et al. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1665–1680.

Rao, S.S.P., Huang, S.C., Glenn St Hilaire, B., Engreitz, J.M., Perez, E.M., Kieffer-Kwon, K.R., Sanborn, A.L., Johnstone, S.E., Bascom, G.D., Bochkov, I.D., et al. (2017). Cohesin Loss Eliminates All Loop Domains. Cell *171*, 305–320.e24.

Remeseiro, S., Cuadrado, A., Carretero, M., Marťnez, P., Drosopoulos, W.C., Cañamero, M., Schildkraut, C.L., Blasco, M.A., and Losada, A. (2012a). Cohesin-SA1 deficiency drives aneuploidy and tumourigenesis in mice due to impaired replication of telomeres. EMBO J. *31*, 2076–2089.

Remeseiro, S., Cuadrado, A., Gómez-López, G., Pisano, D.G., and Losada, A. (2012b). A unique role of cohesin-SA1 in gene regulation and development. EMBO J. *31*, 2090–2102.

Ricci, M.A., Manzo, C., García-Parajo, M.F., Lakadamyali, M., and Cosma, M.P. (2015). Chromatin fibers are formed by heterogeneous groups of nucleosomes in vivo. Cell *160*, 1145–1158.

Rollins, R.A., Korom, M., Aulner, N., Martens, A., and Dorsett, D. (2004). Drosophila Nipped-B Protein Supports Sister Chromatid Cohesion and Opposes the Stromalin/Scc3 Cohesion Factor To Facilitate Long-Range Activation of the cut Gene. Mol. Cell. Biol. *24*, 3100–3111.

Rollins, R.A., Morcillo, P., and Dorsett, D. (1999). Nipped-B, a Drosophila homologue of chromosomal adherins, participates in activation by remote enhancers in the cut and Ultrabithorax genes. Genetics *152*, 577–593.

Sanborn, A.L., Rao, S.S.P., Huang, S.-C., Durand, N.C., Huntley, M.H., Jewett, A.I., Bochkov, I.D., Chinnappan, D., Cutkosky, A., Li, J., et al. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. Proc. Natl. Acad. Sci. *112*, E6456–E6465.

Schaaf, C.A., Kwak, H., Koenig, A., Misulovin, Z., Gohara, D.W., Watson, A., Zhou, Y., Lis, J.T., and Dorsett, D. (2013). Genome-Wide Control of RNA Polymerase II Activity by Cohesin. PLoS Genet. *9*.

Schmidt, D., Schwalie, P.C., Ross-Innes, C.S., Hurtado, A., Brown, G.D., Carroll, J.S., Flicek, P., and Odom, D.T. (2010). A CTCF-independent role for cohesin in tissue-specific transcription. Genome Research. *20*, 578-588.

Schoenfelder, S., Furlan-Magaril, M., Mifsud, B., Tavares-Cadete, F., Sugar, R., Javierre, B.M., Nagano, T., Katsman, Y., Sakthidevi, M., Wingett, S.W., et al. (2015). The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. Genome Res. *25*, 582–597.

Schuldiner, O., Berdnik, D., Levy, J.M., Wu, J.S., Luginbuhl, D., Gontang, A.C., and Luo, L. (2008). piggyBac-Based Mosaic Screen Identifies a Postmitotic Function for Cohesin in Regulating Developmental Axon Pruning. Dev. Cell *14*, 227–238.

Schwarzer, W., Abdennur, N., Goloborodko, A., Pekowska, A., Fudenberg, G., Loe-Mie, Y., Fonseca, N.A., Huber, W., Haering, C.H., Mirny, L., et al. (2017). Two independent modes of chromatin organization revealed by cohesin removal. Nature *551*, 51–56.

Seitan, V.C., Hao, B., Tachibana-Konwalski, K., Lavagnolli, T., Mira-Bontenbal, H., Brown, K.E., Teng, G., Carroll, T., Terry, A., Horan, K., et al. (2011). A role for cohesin in T-cell-receptor rearrangement and thymocyte differentiation. Nature *476*, 467–473.

Serra, F., Baù, D., Goodstadt, M., Castillo, D., Filion, G., and Marti-Renom, M.A. (2017). Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. PLoS Comput. Biol. *13*.

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., and Cavalli, G. (2012). Three-dimensional folding and functional organization principles of the Drosophila genome. Cell *148*, 458–472.

Shen, C.H., Kim, S.H., Trousil, S., Frederick, D.T., Piris, A., Yuan, P., Cai, L., Gu, L., Li, M., Lee, J.H., et al. (2016). Loss of cohesin complex components STAG2 or STAG3 confers resistance to BRAF inhibition in melanoma. Nat. Med. *22*, 1056–1061.

Shintomi, K., and Hirano, T. (2010). Sister chromatid resolution: A cohesin releasing network and beyond. Chromosoma *119*, 459–467.

Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., De Wit, E., Van Steensel, B., and De Laat, W. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nat. Genet. *38*, 1348–1354.

Smemo, S., Tena, J.J., Kim, K.H., Gamazon, E.R., Sakabe, N.J., Gómez-Marín, C., Aneas, I., Credidio, F.L., Sobreira, D.R., Wasserman, N.F., et al. (2014). Obesity-associated variants within FTO form long-range functional connections with IRX3. Nature *507*, 371–375.

Sofueva, S., Yaffe, E., Chan, W.C., Georgopoulou, D., Vietri Rudan, M., Mira-Bontenbal, H., Pollard, S.M., Schroth, G.P., Tanay, A., and Hadjur, S. (2013). Cohesin-mediated interactions organize chromosomal domain architecture. EMBO J. *32*, 3119–3129.

Solomon, D.A., Kim, T., Diaz-Martinez, L.A., Fair, J., Elkahloun, A.G., Harris, B.T., Toretsky, J.A., Rosenberg, S.A., Shukla, N., Ladanyi, M., et al. (2011). Mutational inactivation of STAG2 causes aneuploidy in human cancer. Science. *333*, 1039–1043.

Solomon, D.A., Kim, J.S., Bondaruk, J., Shariat, S.F., Wang, Z.F., Elkahloun, A.G., Ozawa, T., Gerard, J., Zhuang, D., Zhang, S., et al. (2013). Frequent truncating mutations of STAG2 in bladder cancer. Nat. Genet. *45*, 1428–1430.

Splinter, E., Heath, H., Kooren, J., Palstra, R.J., Klous, P., Grosveld, F., Galjart, N., and De Laat, W. (2006). CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. Genes Dev. *20*, 2349–2354.

Stadhouders, R., Vidal, E., Serra, F., Di Stefano, B., Le Dily, F., Quilez, J., Gomez, A., Collombet, S., Berenguer, C., Cuartero, Y., et al. (2018). Transcription factors orchestrate dynamic interplay between genome topology and gene regulation during cell reprogramming. Nat. Genet. *50*, 238–249.

Stigler, J., Çamdere, G., Koshland, D.E., and Greene, E.C. (2016). Single-Molecule Imaging Reveals a Collapsed Conformational State for DNA-Bound Cohesin. Cell Rep. *15*, 988–998.

Sumara, I., Vorlaufer, E., Gieffers, C., Peters, B.H., and Peters, J.M. (2000). Characterization of vertebrate cohesin complexes and their regulation in prophase. J. Cell Biol. *151*, 749–761.

Subramanian, A., Kuehn, H., Gould, J., Tamayo, P., and Mesirov, J.P. (2007). GSEA-P: a desktop application for Gene Set Enrichment Analysis. Bioinformatics *23*, 3251–3253.

Taverna, S.D., Li, H., Ruthenburg, A.J., Allis, C.D., and Patel, D.J. (2007). How chromatin-binding modules interpret histone modifications: Lessons from professional pocket pickers. Nat. Struct. Mol. Biol. *14*, 1025–1040.

Taylor, C.F., Platt, F.M., Hurst, C.D., Thygesen, H.H., and Knowles, M.A. (2014). Frequent inactivating mutations of STAG2 in bladder cancer are associated with low tumour grade and stage and inversely related to chromosomal copy number changes. Hum. Mol. Genet. *23*, 1964–1974.

Tedeschi, A., Wutz, G., Huet, S., Jaritz, M., Wuensche, A., Schirghuber, E., Davidson, I.F., Tang, W., Cisneros, D.A., Bhaskara, V., et al. (2013). WAPL is an essential regulator of chromatin structure and chromosome segregation. Nature *501*, 564–568.

Thol, F., Bollin, R., Gehlhaar, M., Walter, C., Dugas, M., Suchanek, K.J., Kirchner, A., Huang, L., Chaturvedi, A., Wichmann, M., et al. (2014). Mutations in the cohesin complex in acute myeloid leukemia: Clinical and prognostic implications. Blood *123*, 914–920.

Thota, S., Viny, A.D., Makishima, H., Spitzer, B., Radivoyevitch, T., Przychodzen, B., Sekeres, M.A., Levine, R.L., and Maciejewski, J.P. (2014). Genetic alterations of the cohesin complex genes in myeloid malignancies. Blood *124*, 1790–1798.

Tirode, F., Surdez, D., Ma, X., Parker, M., Le Deley, M.C., Bahrami, A., Zhang, Z., Lapouble, E., Grossetete-Lalami, S., Rusch, M., et al. (2014). Genomic landscape of ewing sarcoma defines an aggressive subtype with co-association of STAG2 and TP53 mutations. Cancer Discov. *4*, 1342–1353.

van der Lelij, P., Lieb, S., Jude, J., Wutz, G., Santos, C.P., Falkenberg, K., Schlattl, A., Ban, J., Schwentner, R., Hoffmann, T., et al. (2017). Synthetic lethality between the cohesin subunits STAG1 and STAG2 in diverse cancer contexts. Elife *6*, 1–15.

Vega, H., Waisfisz, Q., Gordillo, M., Sakai, N., Yanagihara, I., Yamada, M., Van Gosliga, D., Kayserili, H., Xu, C., Ozono, K., et al. (2005). Roberts syndrome is caused by mutations in ESCO2, a human homolog of yeast ECO1 that is essential for the establishment of sister chromatid cohesion. Nat. Genet. *37*, 468–470.

Viny, A.D., and Levine, R.L. (2018). Cohesin mutations in myeloid malignancies made simple. Curr. Opin. Hematol. *25*, 61–66.

Viny, A.D., Ott, C.J., Spitzer, B., Rivas, M., Meydan, C., Papalexi, E., Yelin, D., Shank, K., Reyes, J., Chiu, A., et al. (2015). Dose-dependent role of the cohesin complex in normal and malignant hematopoiesis. J. Exp. Med. *212*, 1819–1832.

Waizenegger, I.C., Hauf, S., Meinke, A., and Peters, J.-M. (2000). Two Distinct Pathways Remove Mammalian Cohesin from Chromosome Arms in Prophase and from Centromeres in Anaphase. Cell *103*, 399–410.

Wani, A.H., Boettiger, A.N., Schorderet, P., Ergun, A., Munger, C., Sadreyev, R.I., Zhuang, X., Kingston, R.E., and Francis, N.J. (2016). Chromatin topology is coupled to Polycomb group protein subnuclear organization. Nat. Commun. *7*.

Welch, J.S., Ley, T.J., Link, D.C., Miller, C.A., Larson, D.E., Koboldt, D.C., Wartman, L.D., Lamprecht, T.L., Liu, F., Xia, J., et al. (2012). The origin and evolution of mutations in acute myeloid leukemia. Cell *150*, 264–278.

Wendt, K.S., Yoshida, K., Itoh, T., Bando, M., Koch, B., Schirghuber, E., Tsutsumi, S., Nagae, G., Ishihara, K., Mishiro, T., et al. (2008). Cohesin mediates transcriptional insulation by CCCTC-binding factor. Nature *451*, 796–801.

Whelan, G., Kreidl, E., Wutz, G., Egner, A., Peters, J.M., and Eichele, G. (2012). Cohesin acetyltransferase Esco2 is a cell viability factor and is required for cohesion in pericentric heterochromatin. EMBO J. *31*, 71–82.

Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., and Young, R.A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell 153, 307–319.

Wilson, N.K., Foster, S.D., Wang, X., Knezevic, K., Schütte, J., Kaimakis, P., Chilarska, P.M., Kinston, S., Ouwehand, W.H., Dzierzak, E., et al. (2010). Combinatorial transcriptional control in blood stem/progenitor cells: Genome-wide analysis of ten major transcriptional regulators. Cell Stem Cell *7*, 532–544.

Wutz, G., Varnai, C., Nagasaka, K., Cisneros, D.A., Stocsits, R., Tang, W., Schoenfelder, S., Jessberger, G., Muhar, M., Hossain, J.M., et al. (2017). CTCF, WAPL and Pds5 proteins control the formation of TADs and loops by cohesin. EMBO J. 36, 3573-3599.

Xiao, T., Wallace, J., and Felsenfeld, G. (2011). Specific Sites in the C Terminus of CTCF Interact with the SA2 Subunit of the Cohesin Complex and Are Required for Cohesin-Dependent Insulation Activity. Mol. Cell. Biol. *31*, 2174–2183.

Xu, Y., Guo, W., Li, P., Zhang, Y., Zhao, M., Fan, Z., Zhao, Z., and Yan, J. (2016). Long-Range Chromosome Interactions Mediated by Cohesin Shape Circadian Gene Expression. PLoS Genet. *12*.

Yingjun, X., Wen, T., Yujian, L., Lingling, X., Huimin, H., Qun, F., and Junhong, C. (2015). Microduplication of chromosome Xq25 encompassing STAG2 gene in a boy with intellectual disability. Eur. J. Med. Genet. *58*, 116–121.

Yusufzai TM, Tagami H, Nakatani Y, Felsenfeld G. CTCF tethers an insulator to subnuclear sites, suggesting shared insulator mechanisms across species. Mol Cell 2004; 13: 291–298.

Yoshida, K., Sanada, M., Shiraishi, Y., Nowak, D., Nagata, Y., Yamamoto, R., Sato, Y., Sato-Otsubo, A., Kon, A., Nagasaki, M., et al. (2011). Frequent pathway mutations of splicing machinery in myelodysplasia. Nature *478*, 64–69.

Zhang, N., Kuznetsov, S.G., Sharan, S.K., Li, K., Rao, P.H., and Pati, D. (2008). A handcuff model for the cohesin complex. J. Cell Biol. *183*, 1019–1031.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nussbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*.

Zhang, Y., McCord, R.P., Ho, Y.J., Lajoie, B.R., Hildebrand, D.G., Simon, A.C., Becker, M.S., Alt, F.W., and Dekker, J. (2012). Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. Cell *148*, 908–921.

Zuin, J., Dixon, J.R., van der Reijden, M.I.J.A., Ye, Z., Kolovos, P., Brouwer, R.W.W., van de Corput, M.P.C., van de Werken, H.J.G., Knoch, T.A., van IJcken, W.F.J., et al. (2014). Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. Proc. Natl. Acad. Sci. U. S. A. *111*, 996–1001.

ANNEX

Table 1. RT-qPCR primer sequences

RT-qPCR primers

name	sequence (5'-3')
S100A9 Fw	TCATCAACACCTTCCACCAA
S100A9 Rev	GTGTCCAGGTCCTCCATGAT
S100A12 Fw	GAGCTTGCAAACACCATCAA
S100A12 Rev	CAATGGCTACCAGGGATATGA
S100A8 Fw	ATGCCGTCTACAGGGATGAC
S100A8 Rev	ACGCCCATCTTTATCACCAG
S100A6 Fw	AAGCTGCAGGATGCTGAAAT
S100A6 Rev	CCCTTGAGGGCTTCATTGTA
S100A4 Fw	GATGAGCAACTTGGACAGCA
S100A4 Rev	CTTCCTGGGCTGCTTATCTG
S100A3 Fw	GTGCACCTTCCAGGAATACG
S100A3 Rev	ACATTCCCGAAACTCAGTCG
S100A2 Fw	AGCTTTGTGGGGGAGAAAGT
S100A2 Rev	CAGTGATGAGTGCCAGGAAA
S100A16 Fw	GCGAGATGCTCCAGAAAGAG
S100A16 Rev	CCGCCTATCAAGGTCCAGTA
S100A14 Fw	CTGACCCCTTCTGAGCTACG
S100A14 Rev	TCCAATCAGCTCCCAGAAAC
GAPDH Fw	TGCACCACCAACTGCTTAGC
GAPDH Rev	GAGGGGCCATCCACAGTCTTC

Table 2. ChIP-qPCR primer sequences

ChIP-qPCR primers

name	sequence (5'-3')	coordinates
c1 Fw	TCCCTTCCACTTCCTTCCTT	chr1:153325179-153325324
c1 Rev	TGAATGACTGGCTTGGTTTG	
c2 Fw	CCTTTTCCCTCCTTCCTCTG	chr1:153467509-153467630
c2 Rev	TAGAAAGTGGGTGGGACCAG	
c3 Fw	AGACCACGACGATTGGAGAC	chr1:153483538-153483657
c3 Rev	CCCTGGACACTGGCATCTAC	
c4 Fw	TCTTCCCAGAAACACCCAAC	chr1:153545317-153545448
c4 Rev	AATCACCACCCACTCTCCTG	
c5 Fw	GCCAACCAATGAATGATGTG	chr8:127711139-127711599
c5 Rev	GAGGAGCTACAGCCAACAGG	
c6 Fw	TAGGTGCAACCTTTGGCTTC	chr12:69751204-69751671
c6 Rev	TCCGCTACCTTGACTCATCC	
c7 Fw	TGTGTGACCTCAGGCAAGTT	chr17:36823672-36824158
c7 Rev	ACGTGGTCTCTGCACACAAG	
c8 Fw	AAGATCTGGAAGGGGTGCTT	chr8:102313456-102314020
c8 Rev	CACACACTCACACCCACACA	
o1 Fw	TGCTCACCTGTGAAGCAATC	chr1:153330234-153330371
o1 Rev	GGCAGCTCACTTACCAAAGC	
o2 Fw	ATGCACTTACCCCTCAATGC	chr1:153348018-153348139
o2 Rev	TTCACCCAAGGGCTAAGATG	
o3 Fw	CACACACCCCAATAATGCTG	chr1:153507886-153508015
o3 Rev	CTTGGCCCTAGTGTGCTCTC	
o4 Fw	GCTCTGGGCAGTGAACATTT	chr1:153517468-153517588
o4 Rev	GACTGAGTGAGGGGTGGAAA	
o5 Fw	GTGGGGAGATAGTGGCTGAG	chr1:153585586-153585708
o5 Rev	GTGCCAGTTTTCACCACTCC	
negative 1 Fw	CCGTGAGGCATGGTTTACTT	chr3:177419867-177420010
negative 1 Rev	CAAGAGATGGCCAAACCAAT	
negative 2 Fw	CATAGGGTCAGCAAGGGAGA	chr1:177636432-177636552
negative 2 Rev	CCTAAAACAGCACCCTCCAG	

Table 3. Datasets

Type of data	Sample Name	read counts >=Q30	Called peaks	Source
ChIP-seq	INPUT HMEC	32277764		this study
ChIP-seq	SA1 HMEC	97741618	45,985	this study
ChIP-seq	SA2 HMEC	98226569	82,077	this study
ChIP-seq	SMC1 HMEC	95778603	68,673	this study
ChIP-seq	CTCF HMEC	17641580	57,776	*Wang et al. (2012) / **GSE30263**
ChIP-seq	INPUT HMEC	14808838		*
ChIP-seq	INPUT MCF10A	24028387		this study
ChIP-seq	SA1 MCF10A	61973461	22,510	this study
ChIP-seq	SA2 MCF10A	73440077	37,308	this study
ChIP-seq	SMC1 MCF10A r1	64699651	31,976	this study
ChIP-seq	SMC1 MCF10A r2	27303617	45,530	this study
ChIP-seq	SMC1 MCF10A merged	92003268	50,229	this study
ChIP-seq	CTCF MCF10A	30115216	50,084	* Ross-Innes et al. (2011) / **E-MTAB-740**
ChIP-seq	INPUT MCF10A	19458675		*
ChIP-seq	Zmym2 MCF10A	45156594	1,017	this study
ChIP-seq	YAP1-5SA MCF10A		9,393	* Jang et al (2017) / **GSE97972**
ChIP-seq	INPUT HCAEC	66200000		this study
ChIP-seq	SA1 HCAEC	73350000	40,796	this study
ChIP-seq	SA2 HCAEC	67000000	39,072	this study
ChIP-seq	SMC1 HCAEC	67813000	79,448	this study
ChIP-seq	CTCF HUVEC	23010966	58,845	***Pope et al. (2014) / **GSM822279**
ChIP-seq	SA1 MCF10A siC r1	22086186	11,175	this study
ChIP-seq	SA1 MCF10A siC r2	24942790		this study
ChIP-seq	SA2 MCF10A siC r1	21790656	28,043	this study
ChIP-seq	SA2 MCF10A siC r2	22717313		this study
ChIP-seq	SA1 MCF10A siSA1 r1	23027702	1,353	this study
ChIP-seq	SA1 MCF10A siSA1 r2	20729535		this study
ChIP-seq	SA2 MCF10A siSA1 r1	23870685	32,931	this study
ChIP-seq	SA2 MCF10A siSA1 r2	20374192		this study
ChIP-seq	SA1 MCF10A siSA2 r1	21306063	17,522	this study
ChIP-seq	SA1 MCF10A siSA2 r2	22825770		this study
ChIP-seq	SA2 MCF10A siSA2 r1	20848771	9,831	this study
ChIP-seq	SA2 MCF10A siSA2 r2	28872448		this study
ChIP-seq	INPUT MCF10A siC	31020560		this study
ChIP-seq	INPUT MCF10A siSA1	32254566		this study
ChIP-seq	INPUT MCF10A siSA2	30803519		this study
ChIP-seq	SMC1 HAP1 wt	24054541	15,157	*Haarhuis et al. (2017) / **GSE95015**
ChIP-seq	SMC1 HAP1 Wapl KO	22673617	33,607	*
Re-ChIP-seq	SA2-IgG MCF10A	1028764		
Re-ChIP-seq	SA2-SA1 MCF10A	923652		
Chromatin states	HMEC and HCAEC			Ernst et al. (2011)
Hi-C	HMEC			Rao et al. (2014)
RNA-seq	MCF10A siC r1	21908005		this study
RNA-seq	MCF10A siC r2	21475239		this study
RNA-seq	MCF10A siC r3	22540536		this study
RNA-seq	MCF10A siSA1 r1	22352385		this study
RNA-seq	MCF10A siSA1 r2	21228796		this study
RNA-seq	MCF10A siSA1 r3	23140541		this study
RNA-seq	MCF10A siSA2 r1	21925036		this study
RNA-seq	MCF10A siSA2 r2	20746842		this study
RNA-seq	MCF10A siSA2 r3	21457678		this study
RNA-seq	MCF10A siCTCF r1	21976840		this study
RNA-seq	MCF10A siCTCF r2	23254209		this study

*** Raw data reanalysed according to parameters explained in Methods.**

Table 4. Proteomic analysis of immunoprecipitates obtained with SA1 and SA2 antibodies in MCF10A extract

MCF10A whole cell extract

FUNCTION	Protein names	Gene names	-Log (Student's t-test p-value)		MS/MS Count									
			SA2-IgG	SA1-SA2	IgG_1	IgG_2	SA1_1	SA1_2	SA2_1	SA2_2	IgG total	SA1 total	SA2 total
REGULATION OF TRANSCRIPTION	Zinc finger MYM-type protein 2	ZMYM2	3.69	-3.72	0	0	0	0	5	5	0	0	10
	Transcriptional coactivator YAP1	YAP1	2.78	-2.39	0	0	0	1	9	9	0	1	18
	Aryl hydrocarbon receptor nuclear translocator-like protein 1	ARNTL	2.71	-2.81	0	0	0	0	6	5	0	0	11
	Circadian locomoter output cycles protein kaput	CLOCK	2.32	-2.25	0	0	0	0	1	2	0	0	3
	Zinc finger protein 281	ZNF281	1.97	-1.86	2	1	2	2	11	9	3	4	20
	Forkhead box protein K1	FOXK1	1.27	-0.75	0	0	0	1	3	1	0	1	4

FUNCTION	Protein names	Gene names	-Log (Student's t-test p-value)		MS/MS Count								
			SA1-IgG	SA2-IgG	IgG_1	IgG_2	SA1_1	SA1_2	SA2_1	SA2_2	IgG total	SA1 total	SA2 total
COHESIN subunits	Structural maintenance of chromosomes protein 1A	SMC1A	5.4	8.3	0	0	10	4	41	41	0	14	82
	Structural maintenance of chromosomes protein 3	SMC3	2.3	4.1	3	1	24	28	65	71	2	18	41
	Double-strand-break repair protein rad21 homolog	RAD21	3.1	4.6	1	1	8	10	22	19	4	52	136
	Cohesin subunit SA2	STAG2	2.2	3.2	0	0	5	6	18	25	0	11	43
	Cohesin subunit SA1	STAG1	4.8	1.4	0	0	20	28	0	0	0	48	0
COHESIN associated factors	Sister chromatid cohesion protein PDS5 homolog A	PDS5A	0.6	0.9	0	0	0	1	2	4	0	1	6
	Sister chromatid cohesion protein PDS5 homolog B	PDS5B	0.9	2.4	2	2	6	5	23	25	4	11	48
	Wings apart-like protein homolog	WAPL	0.9	2.9	0	0	3	1	19	24	0	4	43

Table 5. HiC data

Experiment	Reads	Valid int.	%	Dangling	%	Extra DE	%	Duplicates	%	R. breaks	%	Error	%	Self circle	%
SiControl 1	248,647,664	133,420,784	53.7%	12,987,899	5%	39,167,568	16%	90,124,204	36%	168,103	0%	1,998,289	1%	133,821	0%
SiControl 2	227,310,766	151,682,621	66.7%	21,179,856	9%	50,345,583	22%	32,998,689	15%	154,928	0%	1,653,370	1%	75,164	0%
siControl merge	**475,958,430**	**285,103,405**	**59.9%**												
siSA1 1	274,846,524	156,125,223	56.8%	15,092,103	5%	43,726,123	16%	95,116,173	35%	209,301	0%	1,442,891	1%	123,768	0%
siSA1 2	213,454,660	116,036,473	54.4%	26,932,052	13%	46,603,546	22%	33,183,821	16%	187,927	0%	3,251,905	2%	72,204	0%
siSA1 merge	**488,301,184**	**272,161,696**	**55.7%**												
siSA2 1	269,214,001	143,439,516	53.3%	16,920,813	6%	48,370,431	18%	98,431,718	37%	176,815	0%	1,512,800	1%	166,514	0%
siSA2 2	203,103,729	160,391,679	79.0%	12,117,582	6%	39,273,079	19%	16,245,872	8%	149,887	0%	1,195,059	1%	80,968	0%
siSA2 merge	**472,317,730**	**303,831,195**	**64.3%**												